MATEMATIK

NOTAT 15

DESKRIPTIV STATISTIK

AF:

CAND. POLYT.

MICHEL MANDIX

SIDSTE REVISION: APRIL 2024

Oversigt over græske bogstaver:

Kapitaler	Minuskler	Navn
Α	α	Alfa
Γ	γ	Gamma
Ε	ε	Epsilon
Η	η	Eta
Ι	ι	Jota
Λ	λ	Lambda
Ν	ν	Ny
Ο	o	Omikron
Ρ	ρ	Rho
Τ	τ	Tau
Φ	φ	Phi
Ψ	ψ	Psi

Kapitaler	Minuskler	Navn
Β	β	Beta
Δ	δ	Delta
Ζ	ζ	Zeta
Θ	θ	Theta
Κ	κ	Kappa
Μ	μ	My
Ξ	ξ	Xi
Π	π	Pi
Σ	σ	Sigma
Υ	υ	Ypsilon
Χ	χ	Chi
Ω	ω	Omega

Indholdsfortegnelse:

Deskriptiv statistik

Introduktion:

Deskriptiv statistik betyder ”beskrivende statistik”. Bare tænk på engelsk. ”To describe” betyder ”at beskrive”.

Deskriptiv statistik dækker over en stor samling af metoder, der netop beskriver de data, som er skaffet ved en undersøgelse eller dataindsamling.

Således kan deskriptiv statistik IKKE anvendes til at lave forudsigelser om en population ud fra de indsamlede data, da de omtalte metoder kun kan beskrive data og ikke generalisere ud fra data.

Så hvad kan deskriptiv statistik bruges til? Det er et fantastisk værktøj til at få overblik over indsamlede data og kan – i hænderne på de rette – bruges til at præsentere data i artikler etc.

Ordliste:

Observation: Anvendes, når det er noget helt bestemt man undersøger. F.eks. hvilke karakterer, der er givet i en klasse, hvilke skostørrelser der bruges i en virksomhed, hvor mange producerede maskiner der er fejl på etc.
Eksempel:
Handler undersøgelsen om karakterer, kan observationerne være: -3, 00, 02, 4, 7, 10 eller 12.

Hændelse: En hændelse er noget, som sker. Det er helt konkret og det er noget som man registrerer som led i en undersøgelse.
Eksempel:
Det er en hændelse, hvis en elev til eksamen får et 7-tal.

Udfald: Handler en statistik mere om chancebetonede sandsynligheder, taler man om udfald. Det kan være kast med en mønt eller kast med terninger.
Er der tale om kast med en mønt, kan udfaldene være: Plat eller krone.
Er der tale om kast med en terning, kan udfaldene være: en etter, en toer, en treer, en firer, en femmer eller en sekser.

Elementer:

Population:

Stikprøve:

Når man skal lave en statistisk opgave fra bunden, skal man først gøre sig klart om man skal bruge "Diskrete variable" eller "Grupperede variable".

Forskellen på de "**Diskrete variable**" og de "**Grupperede variable**" er følgende:

Diskrete variable bruger man, når man gentagne gange har registreret bestemte størrelser. Tag for eksempel følgende eksempel, hvor man har målt hvor mange gange man har registreret familier med netop 1 barn, 2 børn osv. Bemærk, at en familie ikke her kan have 1,8 børn. Alle registreringer er på NETOP 1, 2, 3 … osv. børn.

Grupperede variable bruger man, hvis de registreringer man har ikke er "præcise". Eller med andre ord – hvis de målinger man foretager, ikke er ens, men derimod godt kan indpasses i et interval. Tag for eksempel følgende eksempel, hvor man godt kan læse lektier i 2,8 timer – også selvom der ikke er nogen præcis værdi på 2,8 timer. Men her kan man indpasse registreringen i intervallet]0;3] timer. (Læses: "Fra 0 til 3 timer".)

Hvis der er 10.000 svarkategorier, vil man nok også tilpasse data i grupperinger, da der er dels er for mange svarkategorier til at de er håndterbare og dels at der (måske) er mange værdier, som ikke er benyttet.

I dette notat vil definitioner være skrevet med blå skrift, og eksempler vil være skrevet med rødt.

Diskrete variable – et eksempel:

Kriteriefunktion: For bedre at kunne forstå definitionerne, betragtes følgende eksempel, som er taget fra
"HTX Mat B2", af Martinus et.al.
Eksemplet er gennemregnet, men desuden gennemgået i Excel, hvor alle funktionerne er forklaret i den rækkefølge de bruges.

I en lille studieretning er der 9 (ni) elever. De har fået karaktererne: 7, 4, 10, 7, 7, 12, 4, 7 og 02 i matematik.

Det bemærkes, at svarene står i den rækkefølge de er indkommet i, og der er derfor tale om et uordnet datasæt.

For bedre at kunne overskue datasættet (og desuden til brug for senere udregninger) kan datasættet ordnes i en mængde, så karaktererne (elementerne i datasættet) stiger numerisk fra venstre mod højre.

Datasættet kaldes for X, og da der er 9 elementer, siger man, at populationen, n, er lig med 9.

$X = \{2,4,4,7,7,7,7,10,12\}$, $n = 9$.

Det vil ofte ikke være muligt at sortere data manuelt. Enten fordi datamængden er for overvældende – altså er der er alt for meget data. Det kunne typisk være fordi data ankommer i en Excel-fil eller lignende. Her giver det ikke mening at sortere data, men man vælger typisk i stedet at benytte sig af CAS til at udregne de forskellige deskriptorer.
Da det er hele klassen (en lille klasse), siger man at n er en population, da det er hele gruppen, som undersøges. Havde det været et udsnit (i så fald ville man nok sige, at data var ugyldige pga. den beskedne mængde), ville man have sagt, at n var størrelsen på en *stikprøve*.

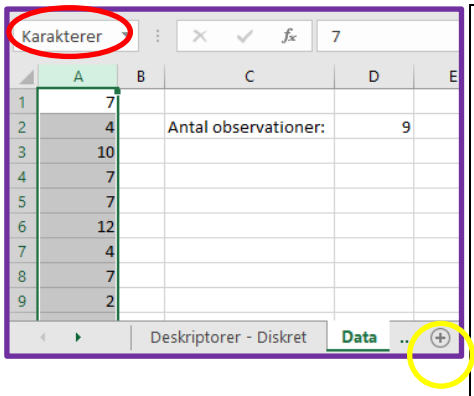
Definition: Elementer:

Elementer dækker over hvert eneste tal i et datasæt. Altså de værdier, som undersøgelsen kan antage. Bemærk, at observationerne ikke kan "falde udenfor". I eksemplet er det f.eks. ikke muligt at give karakteren 6,5. Eftersom der arbejdes med et diskret datasæt, bør elementerne sorteres efter stigende talværdi, skrevet i en liste fra venstre mod højre.

$X = \{x_1, x_2, \ldots, x_n\}$

$X = \{2,4,4,7,7,7,7,10,12\}$, $n = 9$

Det kan være praktisk – særligt hvis der er mange observationer – at registrere antallet af forekomster af hver enkelt slags observation. Antallet af forskellige slags observationer skrives som k.



I Excel, kan man med fordel indtaste data i et andet regneark. Det skal nævnes for en ordens skyld, at det, de fleste mennesker kalder et "regneark" i virkeligheden er en "regnearksmappe" – dvs. et system, som indeholder flere regneark. Et regneark er da blot et faneblad, som vises nederst i Excel-vinduet.
Flere regneark kan tilføjes ved at klikke på det lille plus "+" (markeret med en gul cirkel på figuren) ved siden af de eksisterende regneark.
Regneark er som standard navngivet "Ark 1", "Ark 2", etc., men hvis man dobbeltklikker på navnet, får man mulighed for at omdøbe fanebladet til et andet navn.

Når data er skrevet (uordnet – dvs. i den rækkefølge, som de indkommer), er det en god ide at navngive dataområdet. Det vil spare mange indtastninger og meget besvær, og det er væsentligt mere sikkert at arbejde med navngivne områder fremfor at man skal markere celler hver gang der skal beregnes noget. Det ses her, at celleområdet er blevet navngivet "Karakterer". Det ses (og indtastes) i navneboksen, som er markeret med en rød ellipse på figuren.

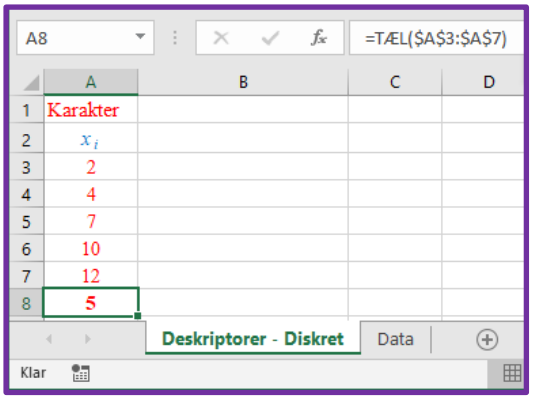
I eksemplet er markeret cellerne $A$1:$A$9, hvilket betyder at navnet "Karakterer" altid vil pege på præcis disse 9 celler. Men det er muligt at "fremtidssikre" sit regneark ved at markere hele kolonnen. I så fald vil markeringen hedde: $A:$A og alle værdier, som skrives i kolonne A vil være med i statistikken.
Det giver således plads til 1.048.576 værdier (en på hver række – måske fratrukket en eller to for overskriften/overskrifterne), så det skulle være rigeligt til langt de fleste undersøgelser.

Man kan lige – i samme omgang, tælle antallet af observationer, *n*. Dette kan gøres med funktionen: "=TÆL(Karakterer)". Dette resultat kan placeres et vilkårligt sted i "Data"-regnearket, men det anbefales at gøre det i en af de øverste rækker, så det er nemt at finde igen. Bare det ikke er i kolonne A. Den er jo reserveret til at kopiere data ind i. Her er det gjort i celle D2.

I dette eksempel vises antal observationer af de forskellige karakterer, som er givet på holdet. Der er ikke givet "-3" og "00", men resten af karakterskalaen er benyttet. Det betyder, at der er registreret 5 forskellige karakterer. Således er $k = 5$ i dette eksempel, da det svarer til 5 forskellige mulige svar.

Det kan være en rigtig god idé at indsætte værdier i en tabel (regneark) for overskuelighedens skyld.

Så hvis det ikke allerede eksisterer, oprettes nu det regneark, som hedder "Deskriptorer – Diskret". Det er her, alle deskriptorerne skal udregnes.



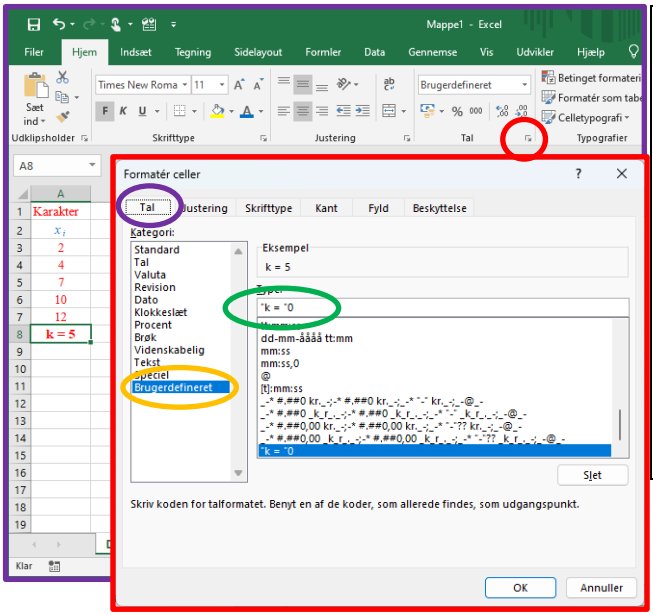
En enkelt manuel udregning er dog smart lige at lave. Det er at tælle antallet af svarkategorier. Det er jo den værdi, der kaldes k.

Dette gøres vha. funktionen "TÆL".

De forskellige karaktermuligheder indtastes og i cellen nedenunder skrives: "=TÆL($A$3:$A$7), hvilket betyder, at der skal tælles antallet af tal i cellerne fra A3 til A7.

Her er det ikke nogen udfordring, men i større undersøgelser er det en praktisk funktion at kende til.

Deskriptiv statistik



Det ser altid godt ud (og er som regel påkrævet), at man adresserer sine udregninger. Med andre ord, at man skriver hvad det er, man udregner.

En måde at gøre det på i Excel er at benytte formatering. Det kan som regel ikke lade sig gøre at kombinere tekst og tal, fordi Excel kun kan regne på rene tal og ikke på tekst. Men man kan snyde på denne måde. Hvis man f.eks. vil have enheder som "cm" eller "kg" efter et tal gøres det på samme måde.

Klik på det lille ikon med pilen i nederste højre hjørne af "Tal"-området. (Her markeret med en **rød** cirkel). Klik derefter på fanebladet "Tal", hvis det ikke allerede er aktivt. (**Lilla** ellipse). Vælg "Brugerdefineret" fra menuen. (**Orange** ellipse). Til sidst skrives i inputlinjen: ' "k = "0 '.
Bemærk, at der skal citationstegn rundt om "k = ". (**Grøn** ellipse).

Definition: Hyppighed:

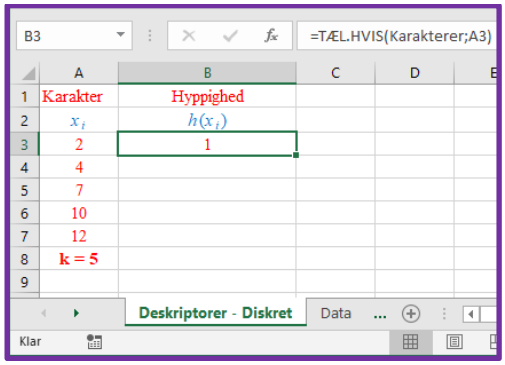
Hyppighed, $h(x_i)$, defineres som antallet af observationer, som "rammer" en bestemt talværdi. Det kan også sige at være det antal gange en bestemt talværdi forekommer blandt elementerne i et datasæt. Eftersom hyppigheden skal beskrives for hvert enkelt element, skrives det: $h(x_i)$, hvor det lille *i* repræsenterer hver enkelt talværdi, som de forskellige elementer kan antage – altså her de forskellige karakterer.

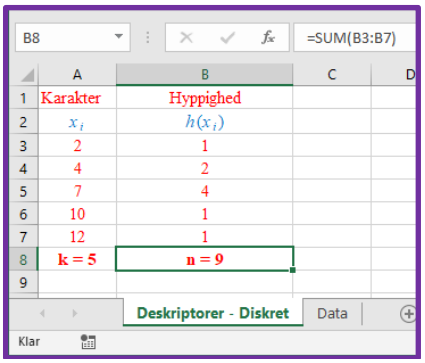
Hvor mange gange optræder en værdi i observationssættet? En simpel optælling. Bemærk, at observationssættet nogle gange kan stå for sig selv, som f.eks. 7, 7, 10, 7, 7, 12, 4, 7, 02, osv. I dette tilfælde må man selv tælle hvor mange gange f.eks. en bestemt karakter er givet. Resultatet af denne optælling, indsættes i skemaet. Andre gange får man bare at vide at karakteren 7 er givet 4 gange, og i dette tilfælde indsætter man ganske enkelt 4 i skemaet udfor karakteren 7.

Hvis man ved lige præcis, hvilke svarkategorier, der kan være tale om og datasættet står uordnet, kan Excel tælle antallet af forekomster.

Det kan, som allerede nævnt, være en fordel at navngive det område, som data står i, men det kan dog også lade sig gøre at markere området, når funktionen indtastes, og det erindres her, at karaktererne i dette eksempel er placeret i et navngivet celleområde, som hedder "Karakterer".

Bruger man Excel, kan man bruge TÆL.HVIS-funktionen. Stå i den celle, hvor resultatet skal stå. Skriv "=TÆL.HVIS(Karakterer;C2)".
Karakterer (= A1:A9) er området med data i. Det kan markeres på mange måder, men det vigtige her er, at tallene er markeret. C2 refererer til den celle, hvor den værdi man vil have optalt står. I dette eksempel, ønskes der at tælle antallet af gange, hvor karakteren 2 er givet.





Som det ses, fastlåses referencen til dataområdet ved at benytte det navngivne område. Ellers går det galt, når formlen kopieres ned til de andre karakterer, som det ses på billedet til højre.
Når formlen trækkes ned, vil den for hver række tælle antallet af forekomster af den karakter, som står i cellen til venstre for disse celler.

Nederst er antallet af observationer lagt sammen. Det er altod en god ide for at se, om det totale antal stemmer.
Det er en helt almindelig SUM-funktion, som derefter er blevet formateret på nøjagtig samme måde, som "k"-værdien blev det.
Det ses, at der – som forventet – er 9 observationer.

Definition: Summeret hyppighed:

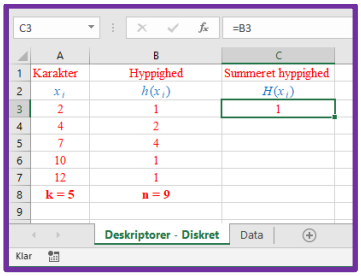
Den summerede hyppighed, H_i, udregnes som hyppigheden for en bestemt kategori, h_i lagt sammen med alle de hyppigheder, som er bestemt for tidligere kategorier – dvs. mindre værdier af i.

Dvs. $H_i = h_i + h_{i-1} + \cdots + h_1$

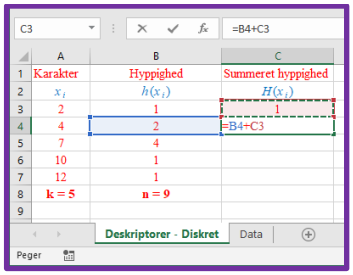
Særligt er $H_1 = h_1$.

$H_2 = h_1 + h_2$, og herfra gælder den generelle formel.

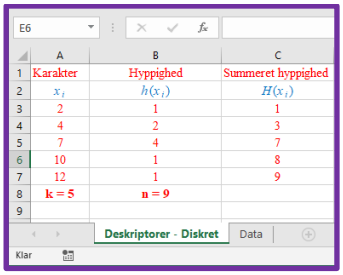
Summerede værdier skal udregnes på to måder. Den første "summerede" hyppighed er jo den første, så den kan jo ikke lægges til en tidligere observeret hyppighed. Så den er jo bare sig selv – eller "lagt sammen med 0".
Resten udregnes som den seneste udregnede summerede hyppighed, lagt sammen med den aktuelle hyppighed. Det ser ud som følger:



Den første summerede hyppighed er "bare" lig med den første hyppighed.



Den anden summerede hyppighed er lig med hyppighed nummer to lagt sammen med den første hyppighed.



Formlen virker, når man kopierer den ned gennem de 5 kategorier.

Deskriptiv statistik

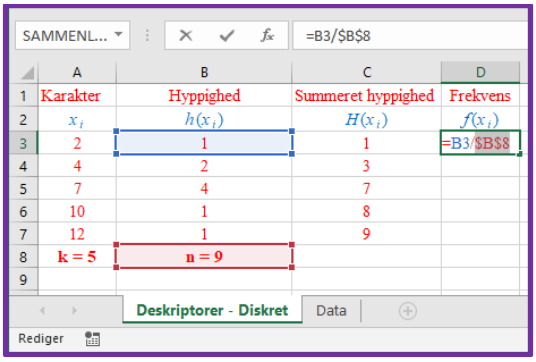
Definition: Frekvens (Den relative hyppighed):

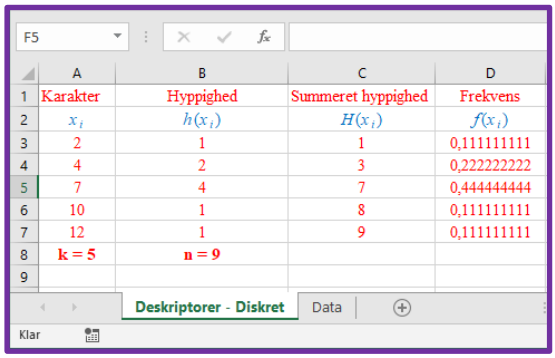
Frekvens (f_i) defineres som hyppigheden i procent af det samlede antal elementer i et datasæt.

$$f_i = \frac{\text{Hyppighed af observationer i en bestemt kategori}}{\text{Population (Antal elementer)}} = \frac{h_i}{n}$$

Frekvensen eller den relative hyppighed. **Leverer data til pindediagrammet**. Udregnes som en bestemt hyppighed divideret med det totale antal observationer. F.eks. for "karakteren 7" gælder: Der er observeret 4 forekomster! 4 divideret med det totale antal observationer (9) = 0,44… Sådan udregnes frekvensen for alle hyppigheder, hver for sig.

Bruger man Excel, bør man oprette en formel! Det gøres smartest i den øverste celle – altså den øverste af de celler, hvor man skal have udregninger. Her skriver man: "=B3/B8". MEN det vil ikke gå godt, fordi når man kopierer denne formel ned, vil alle referencer flytte med. Men da man ALTID (i de 5 udregninger) skal dividere med det totale antal (9) skal referencen til B8 låses. Dette gøres ved at, når man lige har skrevet "B8" i formlen, trykker på "F4"-tasten (på en pc.). Derved kommer der dollartegn på referencen: "=B3/$B$8". Denne formel er korrekt, for når man trækker den ned i fyldhåndtaget, får man det rigtige resultat i alle cellerne.





Man kan evt. afrunde resultatet i Excel til et mindre antal decimaler – afhængigt af opgavens natur. I dette eksempel accepteres standarden i Excel.

Som en kontrol, kan man addere alle frekvenserne. Denne sum SKAL give 1. Det er et udtryk for, at alle adspurgte har svaret med et gyldigt svar. Da der ikke eksisterer tilfælde, hvor eleverne ikke har fået enten 02, 4, 7, 10 eller 12, har det ikke været muligt at svare noget andet. Så summen af alle frekvenserne skal derfor give 1.

Den snedige elev har sikkert allerede bemærket, at frekvensen samtidig er den procentsats, med hvilken en bestemt karakter er givet. Således er karakteren 4 givet 2 gange.

$$f_2 = \frac{h_2}{n} = \frac{2}{9} \approx 0,2222$$

$$\%sats_2 = \frac{2}{9} \cdot 100\,\% \approx 22\,\%$$

Det vil sige, at hvis man lavede præcis den samme undersøgelse igen, så ville der ved hver eneste observation være omkring 22 % chance for, at den givne karakter ville være et 4-tal.

Definition: Summeret frekvens:

Her lægges frekvenserne sammen – på nøjagtig samme måde, som hyppighederne blev summeret i forrige afsnit. Data i denne kolonne kan bruges til at finde ud af, hvor stor en andel er mindre end eller lig med en bestemt værdi? (Det er her, fraktilerne kommer ind i billedet. Mere om det senere,) Denne kolonne forsyner **data til et trappediagram, fraktiler og kvartilsættet**.
Det at "lægge frekvenserne sammen" vil sige, at man hele tiden summerer for hver enkelt observation. Begynd øverst og husk at et trappediagram i denne forbindelse ALTID går opad og går mellem 0 og 1.
Således vil man i eksemplet se, at for den øverste række giver resultatet: 0+0,11=0,11.
Række nr. 2 giver resultatet: (sidste resultat)+(frekvens for næste obs.) = 0,11+0,22=0,33.
Række nr. 3 giver resultatet: (sidste resultat)+(frekvens for næste obs.) = 0,33+0,44=0,77.
Osv…
Bemærk, at resultatet i 3. række (karakteren 7) fortæller os, at i 77 % af de givne karakterer blev der højst givet 7, mens der (aflæses i 2. række) at 33 % af karaktererne var højst 4 osv. Det er derfor, at det er i det summerede trappediagram, at man aflæser fraktiler og kvartiler.

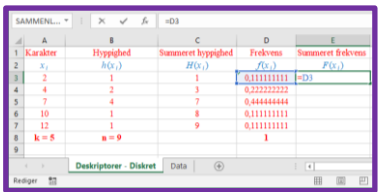
Formlen for den **summerede frekvens** er den "forrige" summerede frekvens, adderet med den "nuværende" frekvens. dvs.:

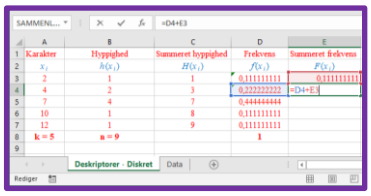
Dvs. $F_i = f_i + f_{i-1} + \cdots + f_1$

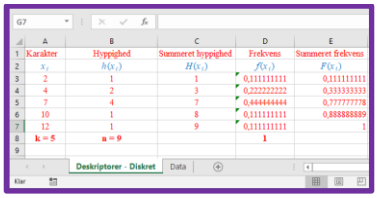
Særligt er $F_1 = f_1$.

$F_2 = f_1 + f_2$, og herfra gælder den generelle formel.

Bruger man Excel, kan man lave en formel. I eksemplet i den øverste række indtastes: "=D3". Det er ikke nødvendigt med mere, da der jo ikke er nogle rækker oven over, som kan bidrage til resultatet.
I række 2 er det lidt mere besværligt. Tast "=D4+E3". Denne formel vil tage den næste frekvens og lægge til det forrige resultat. Denne formel kan trækkes ned i fyldhåndtaget, og vil give det rigtige resultat for hver enkelt række.







Det er alternativt (og marginalt mere besværligt), om man vil udregne den akkumulerede hyppighed og benytte den til at udregne den summerede frekvens. Gør man det, er det blot den akkumulerede hyppighed, som skal divideres med populationen, *n*.
Ordet "Akkumuleret" er præcis det samme som "summeret".

Definition: Produkt:

Produkterne bruges til at finde middelværdien (μ) – også kaldet middeltallet eller gennemsnittet.
Produktet bruges mest til at finde middelværdi (=gennemsnit). Hver række bidrager til produktet med værdien (i dette tilfælde karakteren) ganget med hyppigheden. Dette skyldes, at når f.eks. karakteren 7 er givet 4 gange, så er den samlede "vægt i regnskabet" at der i alt er givet 28 karakterpoints ved hjælp af karakteren 7.

Produktet $x_i \cdot h_i$ giver den enkelte talværdis vægt i kraft af antallet af gange talværdien er blevet observeret.

Deskriptiv statistik

*Bruger man Excel, kan man lave en formel. I eksemplet i den øverste række indtastes: "=A3*B3". Afslut med "Enter"-tasten. Denne formel kan man trække direkte ned i fyldhåndtaget for at lave resten af rækkerne.*

I eksemplet er karakteren (0)2 er givet én gang. I det samlede regnskab giver dette 2 karakterpoints.
2 gange er der givet karakteren 4. Dette bidrager til det samlede regnskab med 2 gange 4 = 8.
etc.
Når man så har fundet produktet for hver værdi, lægges disse sammen. Det vil altså sige, at man har fundet ud af, at der i hele undersøgelsen er givet 60 karakterpoints.

Bruger man Excel, kan man igen bruge SUM-funktionen. Stå i den celle, hvor resultatet skal stå. Tryk på "Sum"-knappen. (Den med ∑ på.) Nu skulle Excel gerne selv vælge cellerne med hyppighederne. Hvis Excel ikke gætter rigtigt, så skal man selv vælge (markere) alle cellerne med hyppighederne.. Afslut med "Enter"-tasten.

Da det vides, at der er 9 elever divideres 60 med 9 for at finde middeltallet $\mu = 6,67$.

Skal man beregne middelværdien, lægger man alle produkterne sammen. Dette tal vil i dette eksempel være den totale sum af karakterpoints, som er givet.

Dette skrives matematisk:

$$x_1 \cdot h_1 + x_2 \cdot h_2 + \ldots + x_n \cdot h_n = \sum_{i=1}^{n} x_i \cdot h_i$$

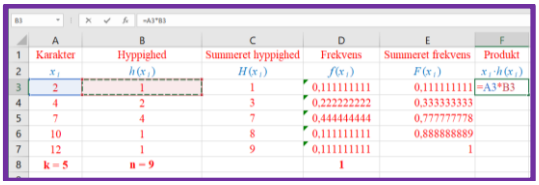
eller i dette eksempel: $\sum_{i=1}^{5} x_i \cdot h_i$, da der er 5 forskellige (anvendte) karakterer.

Middelværdien skrives da som:

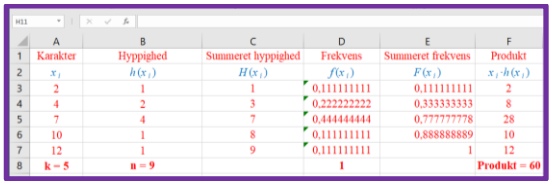
$$\mu = \frac{x_1 \cdot h_1 + x_2 \cdot h_2 + \ldots + x_n \cdot h_n}{n} = \frac{\sum_{i=1}^{n} x_i \cdot h_i}{n}$$

eller i dette eksempel:

$$\mu = \frac{\sum_{i=1}^{5} x_i \cdot h_i}{9} = \frac{2 \cdot 1 + 4 \cdot 2 + 7 \cdot 4 + 10 \cdot 1 + 12 \cdot 1}{9} = \frac{2 + 8 + 28 + 10 + 12}{9} = \frac{60}{9} \approx 6,67.$$



	A	B	C	D	E	F
1	Karakter	Hyppighed	Summeret hyppighed	Frekvens	Summeret frekvens	Produkt
2	x_i	$h(x_i)$	$H(x_i)$	$f(x_i)$	$F(x_i)$	$x_i \cdot h(x_i)$
3	2	1	1	0,111111111	0,111111111	=A3*B3
4	4	2	3	0,222222222	0,333333333	
5	7	4	7	0,444444444	0,777777778	
6	10	1	8	0,111111111	0,888888889	
7	12	1	9	0,111111111	1	
8	k = 5	n = 9		1		



	A	B	C	D	E	F
1	Karakter	Hyppighed	Summeret hyppighed	Frekvens	Summeret frekvens	Produkt
2	x_i	$h(x_i)$	$H(x_i)$	$f(x_i)$	$F(x_i)$	$x_i \cdot h(x_i)$
3	2	1	1	0,111111111	0,111111111	2
4	4	2	3	0,222222222	0,333333333	8
5	7	4	7	0,444444444	0,777777778	28
6	10	1	8	0,111111111	0,888888889	10
7	12	1	9	0,111111111	1	12
8	k = 5	n = 9		1		Produkt = 60

Så, det der er sket her er, at man har taget alle de givne karakterer og lagt sammen. 1 gang er karakteren givet. Det giver et bidrag på 2. Der er givet 2 4-taller, hvilket bidrager med 8, etc. Så i alt er der givet karakterer svarende til 60 karakterpoints.
Men de 60 points blev fordelt på 9 eksaminer, så hvis man tænker, at de har bidraget lige meget, så må gennemsnittet være:

$$\mu = = \frac{60}{9} \approx 6,67.$$

Alternativt kan man beregne middelværdien vha. et andet produkt

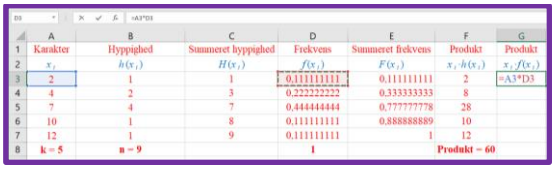
Man kan også finde frem til middeltallet på en anden måde. I stedet for at gange værdien med hyppigheden, lægge alle resultaterne sammen og til sidst dividere resultatet med antallet af observationer (som blev gjort i 5. kolonne), kan man lige så godt tage værdierne og gange med frekvensen. Dette svarer jo til, at man dividerer med antallet af observationer for hver enkelt række med det samme. Dette er den mest benyttede måde – i hvert fald i Excel, fordi man sparer en udregning i den anden ende. Lægger man alle produkterne sammen i 7. kolonne, får man nemlig direkte udregnet middeltallet.

Dette skrives matematisk:

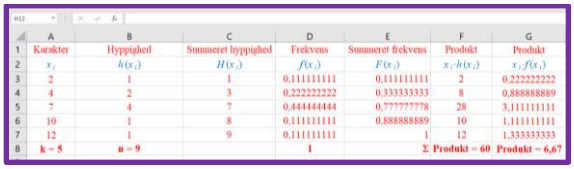
$$\mu = x_1 \cdot f_1 + x_2 \cdot f_2 + \ldots + x_n \cdot f_n = \sum_{i=1}^{n} x_i \cdot f_i$$

eller i dette eksempel:

$$\mu = \sum_{i=1}^{5} x_i \cdot f_i = 2 \cdot 0{,}11 + 4 \cdot 0{,}22 + 7 \cdot 0{,}44 + 10 \cdot 0{,}11 + 12 \cdot 0{,}11 = 0{,}22 + 0{,}88 + 3{,}08 + 1{,}1 + 1{,}32 = 6{,}6$$



	A	B	C	D	E	F	G
1	Karakter	Hyppighed	Summeret hyppighed	Frekvens	Summeret frekvens	Produkt	Produkt
2	x_i	$h(x_i)$	$H(x_i)$	$f(x_i)$	$F(x_i)$	$x_i \cdot h(x_i)$	$x_i \cdot f(x_i)$
3	2	1	1	0,111111111	0,111111111	2	=A3*D3
4	4	2	3	0,222222222	0,333333333	8	
5	7	4	7	0,444444444	0,777777778	28	
6	10	1	8	0,111111111	0,888888889	10	
7	12	1	9	0,111111111	1	12	
8	k = 5	n = 9		1		Produkt = 60	



	A	B	C	D	E	F	G
1	Karakter	Hyppighed	Summeret hyppighed	Frekvens	Summeret frekvens	Produkt	Produkt
2	x_i	$h(x_i)$	$H(x_i)$	$f(x_i)$	$F(x_i)$	$x_i \cdot h(x_i)$	$x_i \cdot f(x_i)$
3	2	1	1	0,111111111	0,111111111	2	0,222222222
4	4	2	3	0,222222222	0,333333333	8	0,888888889
5	7	4	7	0,444444444	0,777777778	28	3,111111111
6	10	1	8	0,111111111	0,888888889	10	1,111111111
7	12	1	9	0,111111111	1	12	1,333333333
8	k = 5	n = 9		1		Σ Produkt = 60	Produkt = 6,67

Således er der lavet en tabel, som giver det fulde overblik over de ting, som skal udregnes:

	De værdier, som undersøgelsen kan antage. I dette tilfælde: Karaktererne	Hvor mange gange optræder en værdi (kat.egori) i observationssættet?		Frekvensen eller den relative hyppighed. Udregnes som en brøkdel af hyppighed divideret med det totale antal observationer. Denne kolonne lægger data til pindediagrammet.	Hvor stor en andel er mindre end eller lig med en bestemt værdi? Denne kolonne lægger data til et trappediagram og kvartilsættet	Produktet bruges mest til at finde middelværdi (= gennemsnit), summen af alle produkterne divideret med antallet af observationer.	Man kan også bruge data i denne kolonne til at finde gennemsnittet. Måske nemmere, da man sparer en udregning
1							
2	Karakter	Hyppighed	Summeret hyppighed	Frekvens	Summeret frekvens	Produkt	Produkt
3	x_i	$h(x_i)$	$H(x_i)$	$f(x_i)$	$F(x_i)$	$x_i \cdot h(x_i)$	$x_i \cdot f(x_i)$
4	2	1	1	0,111111111	0,111111111	2	0,222222222
5	4	2	3	0,222222222	0,333333333	8	0,888888889
6	7	4	7	0,444444444	0,777777778	28	3,111111111
7	10	1	8	0,111111111	0,888888889	10	1,111111111
8	12	1	9	0,111111111	1	12	1,333333333
9	**k = 5**	**n = 9**		**Σ = 1**	**Σ**	**Produkt = 60**	**Produkt = 6,67**

I det følgende beskrives en række deskriptorer som ikke fremgår af det ovenstående skema.

Deskriptiv statistik

Definition: Maksimum:

Maksimum (*Maks*) defineres som den største talværdi, som optræder blandt elementerne i et datasæt.

$X = \{2,4,4,7,7,7,7,10,12\}$, $n = 9$

$Maks = 12$

Idet elementet (12) er den største talværdi blandt de ni elementer, så er *Maks* = 12. Hvis der havde været givet f.eks. 4 12-taller, så ville *Maks* stadig være lig med 12. *Maks* er en absolut værdi.

For fremtidige eksempler, vil de ni celler som indeholder datasættet i Excel være navngivet til "*Datasæt*". Ellers kan man skrive de ni celler som f.eks.: A1:A9. Excel-eksempler er primært konstrueret til at være generiske. Dvs., at i dette (meget) simple eksempel, så vil det virke meget besværligt at sætte så meget i gang for så lidt. Man skal dog tænke på, at datasæt ofte vil være af en ganske anden – og noget større – størrelse.

I Excel:

=MAKS(*Datasæt*)

(På forfatterens computer, virker det IKKE, hvis man bruger MAX).

Denne funktion hedder "MAX …" på engelsk.

Definition: Minimum:

Minimum (Min) defineres som den mindste talværdi, som optræder blandt elementerne i et datasæt.

$X = \{2,4,4,7,7,7,7,10,12\}$, $n = 9$

$Min = 2$

Idet elementet (2) er den mindste talværdi blandt de ni elementer, så er *Min* = 2. Hvis der havde været givet f.eks. 5 02-taller (Bemærk, at der kun regnes med 2. Et foranstillet 0 er ikke væsentligt i matematikken.), så ville *Min* stadig være lig med 2. *Min* er en absolut værdi.

I Excel:

=MIN(*Datasæt*)

Denne funktion hedder "MIN …" på engelsk.

Definition: Middelværdi (μ):

Middelværdien, der skrives som symbolet μ, defineres som gennemsnittet af alle populationens talværdier, som optræder i et datasæt. Er der tale om en stikprøve i stedet for en population, benyttes symbolet $\bar{x}$.

$X = \{x_1, x_2, \ldots, x_n\}$

$$\mu = \frac{\sum_{i=1}^{n} x_n}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$X = \{2,4,4,7,7,7,7,10,12\}$, $n = 9$

$$\mu = \frac{2+4+4+7+7+7+7+10+12}{9} = \frac{60}{9} \approx 6,7$$

Har man allerede lavet skemaet/regnearket, som indeholder observationerne, hyppighederne og frekvenserne, er det allerede beskrevet, hvordan man kan udregne middelværdien.

Middeltallet er dog letpåvirkeligt over for meget små eller meget store værdier. Eksempel kan vise, at hvis der i undersøgelsen pludselig indgår en karakter på -3, så vil middeltallet falde kraftigt, og man kan her vurdere om det er rimeligt, da karakteren -3 ikke er (voldsomt) repræsentativ i datasættet. (Den falder lidt uden for det normale.)

I Excel:
=MIDDEL(*Datasæt*)
Denne funktion hedder "AVERAGE …" på engelsk.

Definition: Variationsbredde:
Variationsbredden defineres som afstanden mellem minimum og maksimum.

$$Variationsbredden = Maks - Min$$

$$Variationsbredden = 12 - 2 = 10$$

I Excel:
=MAKS(*Datasæt*) –MIN(*Datasæt*)

Definition: Median:
Medianen er den værdi, som deler et ordnet observationssæt i to lige store dele. Dvs. hvis alle elementer i et datasæt ordnes efter stigende talværdier, så er medianen det midterste elements talværdi.
Er der et lige antal observationer, så er medianen lig med gennemsnittet af de to midterste elementer.

$X = \{2,4,4,7,7,7,7,10,12\}$, $n = 9$
Da der er et ulige antal observationer (9), må det femte element fra venstre (eller fra højre) være det midterste element.
Værdien af det femte element er 7. Derfor er medianen = 7.

I Excel:
=MEDIAN(*Datasæt*)
Denne funktion hedder "MEDIAN …" på engelsk.

Deskriptiv statistik

Definition: Skævhed:
Hvis middelværdien (μ) er større end medianen, siger man at datasættet er højreskævt.
Hvis middelværdien (μ) er mindre end medianen, siger man at datasættet er venstreskævt.

Middelværdien (μ) er udregnet til at være 6,7.
Medianen er bestemt til at være 7.

Da middelværdien er mindre end medianen, er datasættet venstreskævt. Det betyder samtidig at størstedelen af datasættet ligger til højre for medianen.

I Excel:

Der eksisterer faktisk en funktion i Excel, som hedder "SKÆVHED". Ser man i hjælpefilen for Excel, så er funktionen forklaret som: "Returnerer skævheden for en stokastisk variabel. Skævhed er den grad af asymmetri, der er i en fordeling omkring dens middelværdi. Positiv skævhed indikerer en fordeling med en asymmetrisk hale, der hælder mod mere positive værdier. Negativ skævhed indikerer en fordeling med en asymmetrisk hale, der hælder mod mere negative værdier."

Så det er altså ikke DÉN, man skal bruge.

=SKÆVHED(*Datasæt*)

Til gengæld findes der også funktionen: "SKÆVHED.P", som hjælpefilen beskriver som: "Returnerer skævheden af en distribution baseret på en population: en karakteristik af graden af asymmetri for en distribution omkring dens middelværdi".

=SKÆVHED.P(*Datasæt*)

Her er skævheden beregnet som: $v = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i - \bar{x}^3}{\sigma}$. Denne formel vil naturligvis returnere et tal.
Grundlaget for denne formel er ikke kendt, så konklusionen er, at et tilstrækkeligt resultat vil være at vurdere, om datasættet er højre- eller venstreskævt.

Definition: Typetal:
Typetallet er den eller de observationer, som forekommer oftest (flest gange). Det vil automatisk være det eller de tal, som har den største hyppighed.

$X = \{2,4,4,7,7,7,7,10,12\}$, $n = 9$
Da karakteren 7 har hyppigheden 4, og der ikke er andre tal med den (eller større) hyppighed, så har karakteren 7 den største hyppighed alene.
Derfor er typetallet = 7.

I Excel:
=HYPPIGST(*Datasæt*)
Denne funktion hedder "MODE …" på engelsk.

Definition: Kvartiler:
Kvartiler er nogle ganske bestemte fraktiler (se senere).
Kvartiler opdeler datasættet i 4 (fire) lige store dele.

Der er således 5 kvartiler, som benævnes Q_0, Q_1, Q_2, Q_3 og Q_4.

Q_0 er det samme som datasættets minimum.
Q_2 er det samme som datasættets median.
Q_4 er det samme som datasættets maksimum.

Da maksimum og minimum normalt er underforstået, så er det normalt kun Q_1, Q_2 og Q_3 der skal findes, når det bliver spurgt til kvartilsættet. Medtager man Q_0 og Q_4, kaldes det for "Det udvidede kvartilsæt".

Hvis der er et LIGE antal elementer i datasættet, så beregnes Q_1 som medianen af den første halvdel af datasættet, og Q_3 beregnes som medianen af den sidste halvdel af datasættet.

Hvis der er et ULIGE antal elementer i datasættet, medtages medianen ikke i hverken den første eller sidste halvdel af datasættet, og kvartilerne beregnes som tidligere beskrevet.

(*Det skal nævnes, at der desværre ikke er enighed i verden om, hvordan man beregner kvartilerne, så der kan være andre acceptable metoder, som vil give et anderledes resultat*.)

I eksemplet er der 9 (ni) elementer. Så derfor er der et ulige antal elementer i datasættet.

Første halvdel af datasættet (eksklusiv den midterste værdi, da der er et ulige antal elementer) består af $\{2, 4, 4, 7\}$, og 1. kvartil (Q_1) beregnes derfor som:

$$Q_1 = \frac{4+4}{2} = 4$$

Sidste halvdel af datasættet (eksklusiv den midterste værdi, da der er et ulige antal elementer) består af $\{7, 7, 10, 12\}$, og 3. kvartil (Q_3) beregnes derfor som:

$$Q_3 = \frac{7+10}{2} = 8{,}5$$

Deskriptiv statistik

I Excel:

=KVARTIL.MEDTAG(Matrix;kvartilnummer (0-4))

F.eks. =KVARTIL.MEDTAG(Datasæt;3), hvilket vil returnere tallet 7. I forhold til tidligere udregning er dette ikke korrekt!

=KVARTIL (Matrix;kvartilnummer (0-4))
Denne funktion hedder "QUARTILE …" på engelsk.

F.eks. =KVARTIL (Datasæt;3), hvilket vil returnere tallet 8,5. Dette resultat stemmer i forhold til tidligere udregning, så vil man udregne resultaterne i Excel, anbefales det at benytte =KVARTIL (Matrix;kvartilnummer (0-4)).

Det kan betale sig at lave en lille tabel i Excel, hvor kvartilerne beregnes. Hvilke af dem, som skal inkluderes må være op til den enkelte opgave.

		E	F	G	Formel i kolonne 'G'
		⋮	⋮	⋮	
25	…	0	Q_0 - 0. Kvartil: (Min)	2	=KVARTIL(Datasæt;E25)
26	…	1	Q_1 - 1. Kvartil:	4	=KVARTIL(Datasæt;E26)
27	…	2	Q_2 - 2. Kvartil: (Median)	7	=KVARTIL(Datasæt;E27)
28	…	3	Q_3 - 3. Kvartil:	7	=KVARTIL(Datasæt;E28)
29	…	4	Q_4 - 4. Kvartil: (Maks)	12	=KVARTIL(Datasæt;E29)

Her er ideen med tallene i kolonne 'E', at angive det andet argument til formlen, som er vist længst ude til højre. Da en navngiven reference ALTID er absolut, kan man nøjes med at skrive formlen i celle G2 og derefter trække den ned.

Definition: Kvartilbredde:
Kvartilbredden defineres som afstanden mellem1. kvartil (Q_1) og 3. kvartil (Q_3).

$$Kvartilbredden = Q3 - Q1 = 8,5 - 4 = 4,5$$

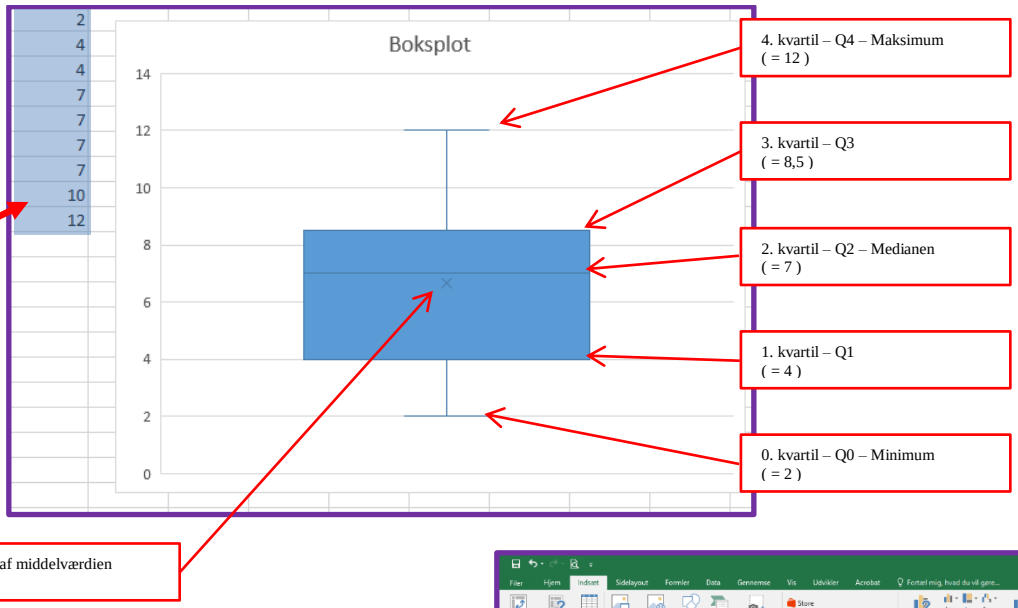
I Excel:

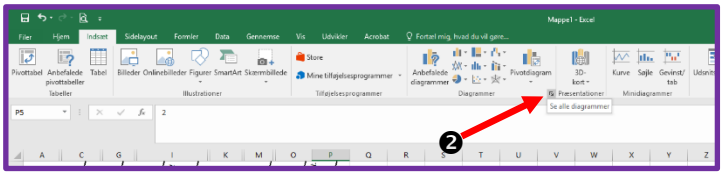
=KVARTIL(*Datasæt;*E28)- KVARTIL(*Datasæt;*E26)

Definition: Boksplot:
Et boksplot er et plot, som viser de fem kvartiler, hvor kvartilbredden er fremhævet.

Da man ofte er interesseret i at få at vide, hvor den midterste halvdel af et datasæt ligger, så er et boksplot en god måde at vise det på. Særligt, hvis man skal sammenligne to datasæt er boksplot en rigtig godt værktøj.

Laver man et boksplot i Excel, indtaster man elementerne i en kolonne, markerer elementerne og indsætter en graf. Denne graf skal være af typen "Kasse med hale, som findes under menuen: "Alle diagrammer".
I modsætning til de fleste andre programmer, vil boksplottet vises lodret i Excel. Det er fint! Det betyder ingenting, hvilken vej det vender.

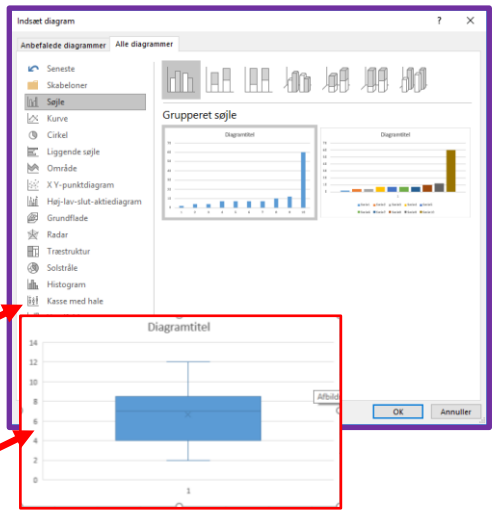


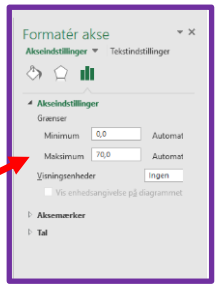


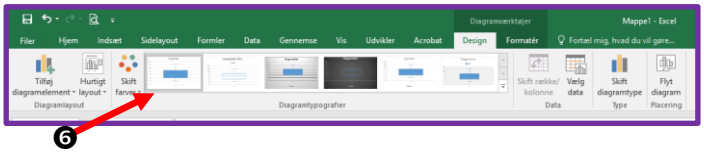
Teksten og tallene indtastes.

- Boksplottet laves således:

1. Marker datasættet.
2. Klik på den lille pil nederst i menufeltet ”Diagram” i fanebladet ”Indsæt”.
3. I fanebladet ”Alle diagrammer” i dialogboksen, vælges ”Kasse med hale”. Der er kun en mulighed. Klik på diagramprototypen og afslut med ENTER.
4. Juster evt. ordinataksen hvis det er nødvendigt. Dobbeltklik på et tal på ordinataksen.
5. I højre sidemenu kan aksernes minimums- og maksimumsværdier indtastes.
6. Når diagrammet er markeret, er menubåndet fokuseret på diagramindstilllinger. Hvis man vil, kan man tilføje bl.a. diagramtitel og mange andre ting i punktet: ”Tilføj diagramelement”.







Definition: Fraktiler:

En fraktil er en procentgrænse, hvor andelen af elementer, som er mindre end eller lig med fraktilen, er mindst lige så stor som fraktilens procentværdi.

Fraktilen beregnes som værdien for x_i, givet: $i \geq \frac{P}{100} \cdot n$,

hvor i er det mindste tal, som opfylder uligheden.

P er fraktilens procent, og n er -

som sædvanlig - populationen, altså antallet af elementer i datasættet.

(*Dette er én blandt flere accepterede måder at beregne fraktiler på. De forskellige måder giver forskellige resultater, men har alle det tilfælles, at de bliver mere ens jo flere elementer, der er i et datasæt.*

0,1-fraktilen er altså den værdi, hvor det først sker, at i er større end eller lig med 0,1.
0,8-fraktilen er den værdi hvor i rammer (eller overskrider) 0,8.
Dette kan aflæses i den summerede frekvens i tabellen eller direkte i trappediagrammet (beskrives senere).
I trappediagrammet kan en fraktil findes ved at afmærke fraktilen op ad frekvens-aksen og derfra tegne en vandret streg ud. Det "trappetrin", som den vandrette streg støder på, svarer til den fraktilens værdi.

Der er nogle fraktiler, som er lidt specielle. Det er dem, man benytter oftest, og de har endda specielle navne:

0,0-fraktil	0. kvartil	Minimum
0,25-fraktil	1. kvartil	Nedre kvartil
0,50-fraktil	2. kvartil	Median
0,75-fraktil	3. kvartil	Øvre kvartil
1,0-fraktil	4. kvartil	Maksimum

De tre (midterste) kvartiler kaldes tilsammen for kvartilsættet eller (for alle fem) det udvidede kvartilsæt.

$X = \{2,4,4,7,7,7,7,10,12\}$, $n = 9$

5 % fraktilen bestemmes:

$$i \geq \frac{5\ \%}{100\ \%} \cdot 9 = 0,45 \Rightarrow \underline{i = 1}$$

(Da der skal rundes op til nærmeste hele element,
skal dette skal forstås som det $i = 1$. element - altså x_1.)

5 % fraktilen er altså det samme som værdien af $\underline{\underline{x_1 = 2}}$.

10 % fraktilen bestemmes:

$$i \geq \frac{10\ \%}{100\ \%} \cdot 9 = 0{,}90 \Rightarrow \underline{i = 1}$$

(Da der skal rundes op til nærmeste hele element, skal dette skal forstås som det $i = 1$. element - altså x_1.)

10 % fraktilen er altså det samme som værdien af $\underline{\underline{x_1 = 2}}$.

Dette svarer til, at 10 % af karaktererne, som er givet er 02 eller derunder.

80 % fraktilen bestemmes:

$$i \geq \frac{80\ \%}{100\ \%} \cdot 9 = 7{,}20 \Rightarrow \underline{i = 8}$$

(Da der skal rundes op til nærmeste hele element, skal dette skal forstås som det $i = 8$. element - altså x_8.)

80 % fraktilen er altså det samme som værdien af $\underline{\underline{x_8 = 10}}$.

Dette svarer til, at 80 % af karaktererne, som er givet er 10 eller derunder.

I Excel:

=FRAKTIL.MEDTAG(Matrix;K (0-1))

F.eks. =FRAKTIL.MEDTAG(*Datasæt*;0,1), hvilket vil returnere tallet 3,6. I forhold til tidligere udregning er dette ikke korrekt!

=FRAKTIL.UDELAD(Matrix;kvartilnummer (0-1))

F.eks. =FRAKTIL.UDELAD(*Datasæt*;0,1), hvilket vil returnere tallet 2. Dette resultat stemmer i forhold til tidligere udregning, så vil man udregne resultaterne i Excel, anbefales det at benytte =FRAKTIL.UDELAD(Matrix;kvartilnummer (0-1)).

Det kan betale sig at lave en lille tabel i Excel, hvor fraktilerne beregnes. Hvilke af dem, som skal beregnes må være op til den enkelte opgave.

		E	F	G			
		⋮	⋮	⋮	FRAKTIL	FRAKTIL.MEDTAG	FRAKTIL.UDELAD
25	⋯	0,05	5 % Fraktil	#NUM!	2,8	2,8	#NUM!
26	⋯	0,1	10 % Fraktil	2	3,6	3,6	2
27	⋯	0,8	80 % Fraktil	10	8,2	3,6	10

Det ses, at Excel kan beregne fraktiler med ikke mindre end tre forskellige funktioner. FRAKTIL, FRAKTIL.MEDTAG og FRAKTIL.UDELAD.
Tilsyneladende er der ingen forskel på FRAKTIL og FRAKTIL.MEDTAG, og de giver begge forkerte resultater i forhold til de udregninger, som er foretaget.
Ser man derimod på FRAKTIL.UDELAD, så giver den de rigtige resultater i to ud af tre tilfælde. I den første udregning, får man en fejlmeddelelse: #NUM. Denne fejl opstår fordi Excel ikke kan interpolere resultatet ud fra de data man har. Typisk vil denne fejl (i denne forbindelse) opstå i de ekstreme yderpunkter – altså nær ved 0 % og ved 100 %.
Denne funktion hedder "PERCENTILE …" på engelsk.

Deskriptiv statistik

Definition: Varians:
For en **population**, *n*, beregnes variansen som den gennemsnitlige kvadratafvigelse.

$$VAR_p = \frac{1}{n} \cdot \sum_{i=1}^{n} h_i \cdot (x_i - \mu)^2 = \frac{h_1 \cdot (x_1 - \mu)^2 + h_2 \cdot (x_2 - \mu)^2 + \ldots + h_n \cdot (x_n - \mu)^2}{n},$$

hvor μ er populationens middelværdi.

For en **Stikprøve**, beregnes variansen som:

$$VAR_s = \frac{1}{n-1} \cdot \sum_{i=1}^{n} h_i \cdot (x_i - \bar{x})^2 = \frac{h_1 \cdot (x_1 - \bar{x})^2 + h_2 \cdot (x_2 - \bar{x})^2 + \ldots + h_n \cdot (x_n - \bar{x})^2}{n-1},$$

hvor $\bar{x}$ er stikprøvens middelværdi.

Da de enkelte observationer kan være både større end middelværdien og mindre end middelværdien, ville man ikke kunne bruge deres sum som et spredningsmål. Derfor har man valgt at kvadrere (sætte dem i anden potens) afvigelserne og lægge dem sammen.
Dette refereres ofte til som **SAK**, som står for **S**ummen af **A**fvigelsernes **K**vadrater.
Ved at dividere med antallet af elementer i datasættet får man den gennemsnitlige kvadratafvigelse, som kaldes for variansen.
Da man ved en stikprøve ikke kender den sande middelværdi, dividerer man i stedet med et mindre tal, så variansen bliver større.
Ved meget store datasæt på f.eks. 10000 elementer, har det ikke den store indflydelse på variansen om man dividerer med 10000 eller 9999. Det viser sig også at være ok, da stikprøvens middelværdi må formodes at være meget tæt på den sande middelværdi ved så stor en stikprøve.
Heraf følger det, at ved en lille stikprøve, så har det imidlertid en meget stor betydning.

$X=\{2,4,4,7,7,7,7,10,12\}$, $n=9$

Ser man på datasættet som kun beskrivende for den lille studieretnings karakterer i matematik, så beregnes variansen som en population:

$$VAR_p=\frac{h_1\cdot(x_1-\mu)^2+h_1\cdot(x_2-\mu)^2+h_1\cdot(x_3-\mu)^2+h_1\cdot(x_4-\mu)^2+h_1\cdot(x_5-\mu)^2}{n}$$

$$\Updownarrow$$

$$VAR_p=\frac{1\cdot(2-6,67)^2+2\cdot(4-6,67)^2+4\cdot(7-6,67)^2+1\cdot(10-6,67)^2+1\cdot(12-6,67)^2}{9}$$

$$\Updownarrow$$

$$VAR_p=\frac{(-4,67)^2+2\cdot(-2,67)^2+4\cdot(0,33)^2+(3,33)^2+(5,33)^2}{9}$$

$$\Updownarrow$$

$$VAR_p=\frac{21,8089+2\cdot 7,1289+4\cdot 0,1089+11,0889+28,4089}{9}$$

$$\Updownarrow$$

$$VAR_p=\frac{21,8089+14,2578+0,4356+11,0889+28,4089}{9}$$

$$\Updownarrow$$

$$VAR_p=\frac{76,00}{9}$$

$$\Updownarrow$$

$$\underline{\underline{VAR_p=8,4}}$$

I Excel:

=VARIANS.P(Matrix)
Denne funktion hedder "VAR.P …" på engelsk.

=VARIANS.P(*Datasæt*)

Bemærk, at de kvadrerede afvigelser skal multipliceres med hyppigheden af de enkelte elementer.

Betragter man derimod datasættet som en stikprøve, som skal beskrive alle eleverne i klassen og ikke kun den lille studieretning, så kender man ikke den sande middelværdi, og man skal derfor dividere med $9-1=8$ i stedet for 9.

$$VAR_{ps} = \frac{h_1 \cdot (x_1 - \mu)^2 + h_1 \cdot (x_2 - \mu)^2 + h_1 \cdot (x_3 - \mu)^2 + h_1 \cdot (x_4 - \mu)^2 + h_1 \cdot (x_5 - \mu)^2}{n-1}$$

$$\Updownarrow$$

$$VAR_s = \frac{1 \cdot (2 - 6{,}67)^2 + 2 \cdot (4 - 6{,}67)^2 + 4 \cdot (7 - 6{,}67)^2 + 1 \cdot (10 - 6{,}67)^2 + 1 \cdot (12 - 6{,}67)^2}{9-1}$$

$$\Updownarrow$$

$$VAR_s = \frac{(-4{,}67)^2 + 2 \cdot (-2{,}67)^2 + 4 \cdot (0{,}33)^2 + (3{,}33)^2 + (5{,}33)^2}{8}$$

$$\Updownarrow$$

$$VAR_s = \frac{21{,}8089 + 2 \cdot 7{,}1289 + 4 \cdot 0{,}1089 + 11{,}0889 + 28{,}4089}{8}$$

$$\Updownarrow$$

$$VAR_s = \frac{21{,}8089 + 14{,}2578 + 0{,}4356 + 11{,}0889 + 28{,}4089}{8}$$

$$\Updownarrow$$

$$VAR_s = \frac{76{,}00}{8}$$

$$\Updownarrow$$

$$\underline{\underline{VAR_s = 9{,}5}}$$

I Excel:

=VARIANS.S(Matrix) ‘
Denne funktion hedder ”VAR.S …” på engelsk.

=VARIANS.S(*Datasæt*)

Variansen er kvadratet på spredningen. Så hvis man vælger at tage kvadratroden af variansen, er det naturligvis fordi man gerne vil have et spredningsmål med samme enhed som de elementer, som er i datasættet.

Spredningen kaldes ofte for den gennemsnitlige afvigelse.

I statistik benytter man ofte græske bogstaver for de sande værdier. Hvis der beregnes på en hel population, vil de beregnede værdier være sande, og derfor bruges det græske bogstav μ (my), som er det tilsvarende bogstav for det latinske ’m’, som igen er valgt for middelværdi. På samme måde er det græske bogstav σ (sigma), som er det tilsvarende bogstav for det latinske ’s’, som igen er valgt for spredning.

Hvis datasættet er normalfordelt, så vil:

95 % af datasættet vil ligge i intervallet : $\mu \pm 1{,}96 \cdot \sigma$
95,4 % af datasættet vil ligge i intervallet : $\mu \pm 2 \cdot \sigma$
90 % af datasættet vil ligge i intervallet : $\mu \pm 1{,}645 \cdot \sigma$

Hvis datasættet IKKE er normalfordelt, så vil (ifølge Chebyshevs teorem):

Mindst 75 % af datasættet ligge i intervallet : $\mu \pm 2 \cdot \sigma$
Mindst 89 % af datasættet ligge i intervallet : $\mu \pm 3 \cdot \sigma$
Mindst 94 % af datasættet ligge i intervallet : $\mu \pm 4 \cdot \sigma$

(Normalfordelingen er en af de vigtigste sandsynlighedsfordelinger og benævnes også Gaussfordelingen. Den er kontinuert og kan principielt omfatte alle reelle tal. Den er symmetrisk og kan entydigt bestemmes ved observationssættets middelværdi og varians. Normalfordelingen bruges som en "model" af hvordan et stort antal statistiske elementer fordeler sig omkring deres gennemsnit. Hvis man for eksempel måler højden eller vægten af hver enkelt person i en stor, ensartet gruppe af mennesker, vil de fleste ligge omkring et vist gennemsnit, mens meget store eller små personer er mere sjældne. Tegner man resultaterne ind i en graf, med højde eller vægt hen ad den vandrette abscisse-akse og f.eks. procent op ad den lodrette ordinat-akse, får grafen den karakteristiske klokkeformede facon, der kan være mere eller mindre fladtrykt eller sammenpresset. Toppunktet ligger ved hele gruppens gennemsnit, markeret på den vandrette akse, og "bulen" omkring toppunktet svarer til det flertal af de målte personer der ligger tæt på gennemsnittet. På begge sider af dette store midterfelt af "gennemsnitsfolk" falder kurven, som tegn på at jo længere væk man kommer fra gennemsnittet, jo sjældnere støder man på folk med sådan en højde eller vægt. Disse dele af kurven omtaler matematikere ofte som haler (ental: en hale).)

Definition: Spredning (Standardafvigelsen):

Spredningen er kvadratroden af variansen.

For en **population**, beregnes spredningen som :

$\sigma = \sqrt{VAR_p}$ eller $\sigma^2 = VAR_p$.

For en **Stikprøve**, beregnes variansen som:

$s = \sqrt{VAR_s}$ eller $s^2 = VAR_s$.

$X = \{2,4,4,7,7,7,7,10,12\}$, $n = 9$

Igen er der forskel. Betragtes datasættet som en population, fås følgende spredning:

$\sigma = \sqrt{VAR_p} = \sqrt{8,4} \approx 2,9$

I Excel:

=STDAFV.P(Matrix)
Denne funktion hedder "STDEV.P …" på engelsk.

= STDAFV.P(*Datasæt*)

Ser man derimod datasættet som en stikprøve, skrives spredningen således:

$s = \sqrt{VAR_s} = \sqrt{9,5} \approx 3,1$

I Excel:

=STDAFV.S(Matrix)
Denne funktion hedder "STDEV.S …" på engelsk.

= STDAFV.S(*Datasæt*)

Det vil være en rimelig antagelse at karaktererne er normalfordelte.
Idet karaktererne er normalfordelte, vil man kunne forvente at 95,4 % af elementerne ligger i intervallet:

$\mu \pm 2 \cdot \sigma$

I dette eksempel bliver det til:

$6,67 \pm 2 \cdot 2,9,$ hvilket svarer til intervallet: $[0,8;12,5]$.

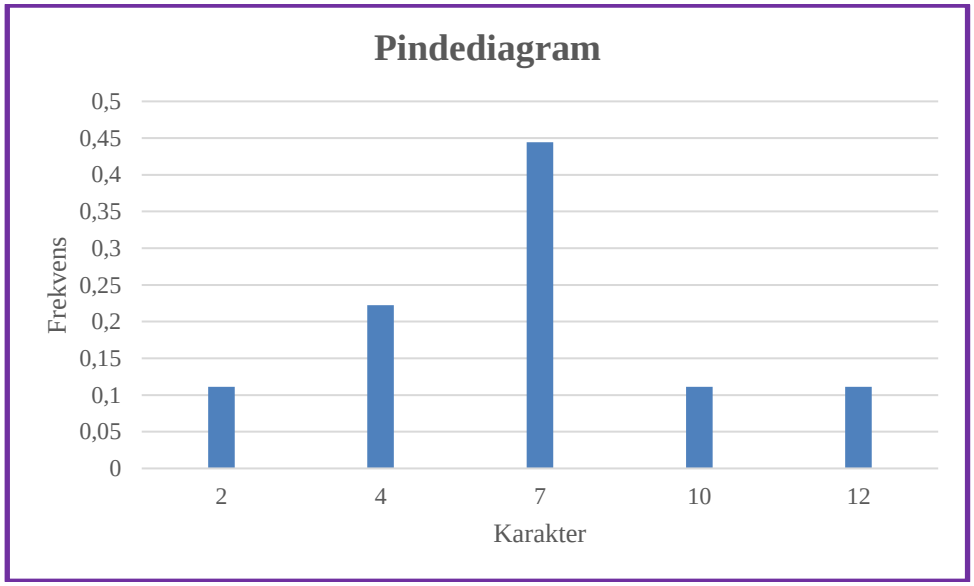
Da både minimum og maksimum er inkluderede i dette interval må hele datasættet ligge i intervallet.

Grunden til at det lige netop er forventeligt at 95,4 % af elementerne ligger i intervallet er, at X er højst 2 standardafvigelser fra μ. Denne forklaring – inklusive bevis – er ikke yderligere beskrevet i dette notat.

Diagrammer for diskrete variable

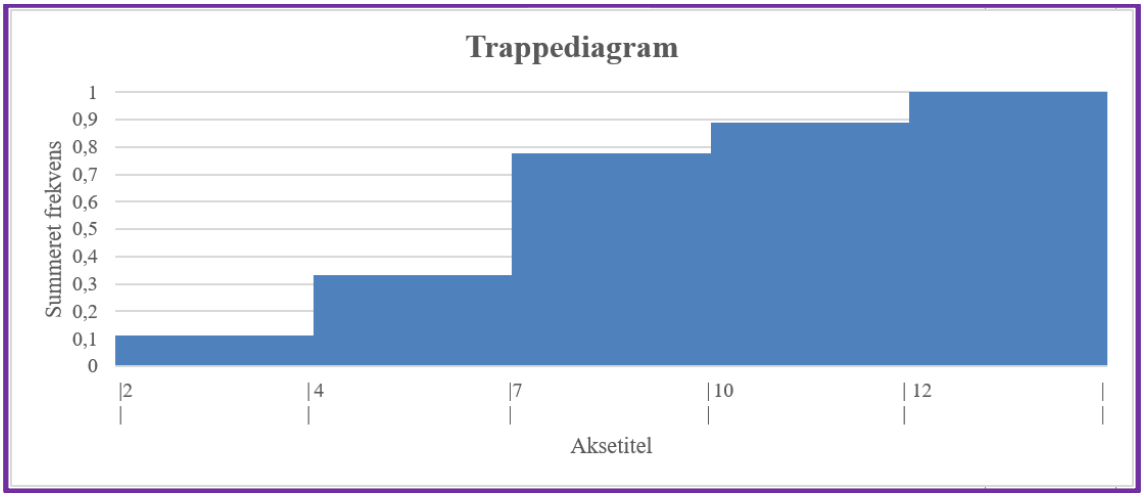
Typisk vælger man at afbilde data for diskrete variable i to diagrammer. Et **pindediagram** og et **trappediagram**.

I pindediagrammet afsætter man værdierne ud af abscisseaksen og frekvensen op ad ordinataksen.
Bemærk at der gælder for BÅDE pindediagrammet og trappediagrammet at ordinataksen ALTID går mellem 0 og 1. (Trappediagrammet går netop fra 0 til 1, medens pindediagrammet – som i nedenstående eksempel – kan være endnu mindre, afhængigt af frekvensernes størrelse.)



(Her er et diagram, som er lavet i Excel. Bemærk, at i egenskaber for dataserie er mellemrummet mellem søjlerne sat til størst muligt (mellemrumsbredde = 500), for at gøre søjlerne så smalle som muligt (pinde). Det er således ikke muligt (I Excel) at lave et helt korrekt pindediagram.

I trappediagrammet nedenunder vises den summerede frekvens. Her afsætter man værdierne ud af abscisseaksen. Sørg for at have den mindste værdi ved ordinataksen. Værdien for den summerede frekvens aftegnes som en vandret streg fra værdiens begyndelse til næste værdi. Her gentages processen for den næste summerede frekvens. Til sidst forbindes de vandrette streger med lodrette streger, så det kommer til at ligne en trappe. (Heraf navnet… ☺) Bemærk desuden, at trappediagrammet altid vil gå opad, da man jo altid lægger sammen. Trappetrinene vil ALTID blive højere og højere, jo længere man kommer hen mod højre. De kan ALDRIG blive lavere!



Her er et trappediagram, som er lavet i Excel. Bemærk, at modsat pindediagrammet, så er mellemrummet mellem søjlerne sat til 0, for at få søjlerne til at støde op mod hinanden.
Det er lidt en tilsnigelse at kalde det et trappediagram, da alle de lodrette streger går hele vejen ned mellem søjlerne. I Excel kan man sagtens fjerne stregerne, hvis man gerne vil.

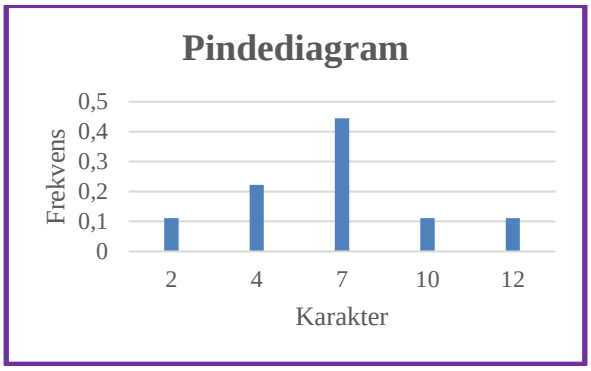
Deskriptiv statistik

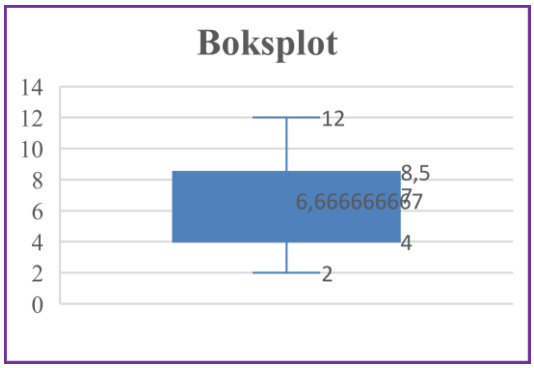
Følgende er en oversigt over de deskriptorer, som beskriver datasættet
$X = \{2,4,4,7,7,7,7,10,12\}$, $n = 9$

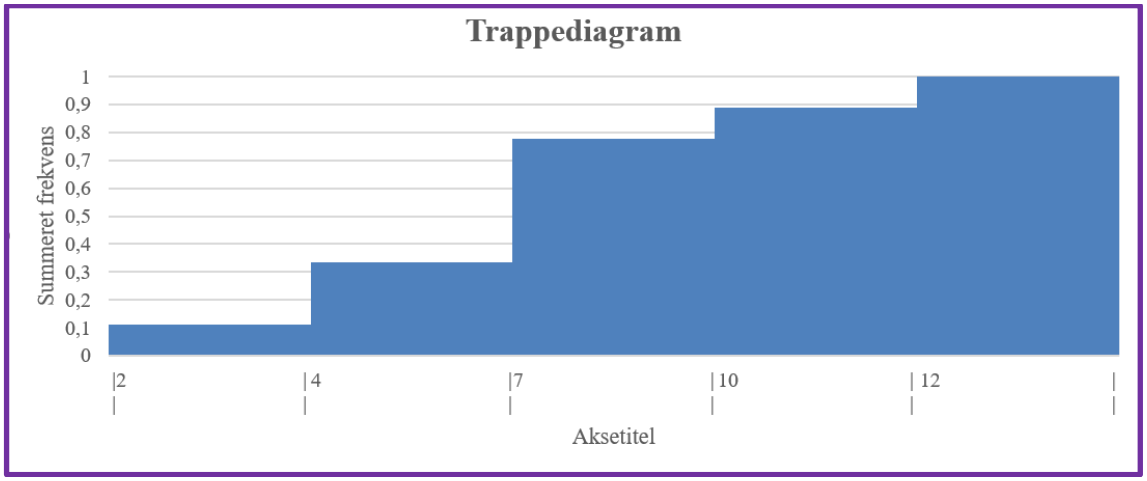
Karakter	Hyppighed	Summeret hyppighed	Frekvens	Summeret frekvens	Produkt	Produkt
x_i	$h(x_i)$	$H(x_i)$	$f(x_i)$	$F(x_i)$	$x_i \cdot h(x_i)$	$x_i \cdot f(x_i)$
2	1	1	0,111111111	0,111111111	2	0,222222222
4	2	3	0,222222222	0,333333333	8	0,888888889
7	4	7	0,444444444	0,777777778	28	3,111111111
10	1	8	0,111111111	0,888888889	10	1,111111111
12	1	9	0,111111111	1	12	1,333333333
k = 5	**n = 9**		**Σ = 1**	**Σ**	**Produkt = 60**	**Produkt = 6,67**

	Populationen:	**9**	Minimum:	**Q0 = Min = 2**
	Middelværdi:	**m = 6,6667**	1. kvartil:	**Q1 = 4**
	Varitionsbredde:	**10**	2. kvartil (Median):	**Q2 = Medianen = 7**
	Typetal:	**7**	3. kvartil:	**Q3 = 7**
	Varians:	**VARp = 8,44**	Maksimum:	**Q4 = Maks = 12**
	Spredning	**s = 2,91**	Kvartilbredde:	**3**
	Skævhed:	**Datasættet er venstreskævt.**	5 % fraktil:	**2**
			10 % fraktil:	**2**
			80 % fraktil:	**10**

Desuden vises de tre grafiske fremstillinger:







Grupperede variable

Som nævnt i indledningen, er det ofte, at de indsamlede data er så forskellige, at man inddeler dem i grupper eller intervaller. I modsætning til de diskrete variable, kan man her tilgodese alle værdier og ikke kun enkelte værdier.

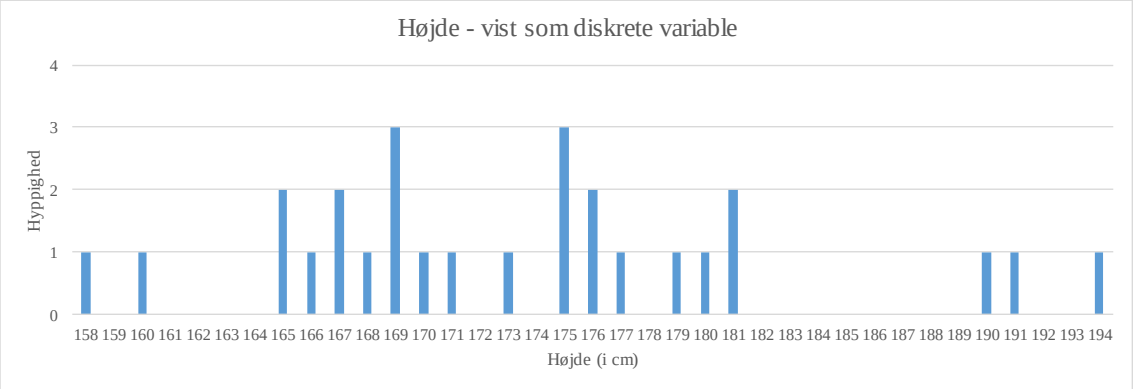
De grupperede variable vil blive forklaret vha. et eksempel.
Igen er eksemplet taget fra "HTX Mat B2", af Martinus et.al.

I en klasse med 27 elever er elevernes højder målt (i *cm*). Resultatet af denne måling er som følger – ordnet efter højde.

158, 160, 165, 165, 166, 167, 167, 168, 169, 169, 169, 170, 171, 173, 175, 175, 175, 176, 176, 177, 179, 180, 181, 181, 190, 191 & 195.

Alle de statistiske deskriptorer beregnes på nøjagtig samme måde, som det blev gjort i kapitlet om Diskrete Variable. Der er dog lige et par ting, som skal overvejes nøje, inden man begynder.

Fra og med 158 (*cm*) til og med 195 (*cm*) er der 38 mulige heltalsværdier, så det kan hurtigt blive til en meget stor tabel. Måske ikke noget udregningsproblem med f.eks. Excel, men det er også svært at overskue en så stor tabel. Desuden er der kun 27 elever, så – især hvis man tænker på, at flere elever kan have samme højde – må der være en del højder, som ikke bliver "brugt". Man siger også, at der er nul-forekomster i datasættet.



Så det er tydeligt, at dette ikke er nogen god fremstillingsmåde til den slags data.

Man kan i stedet vælge at gruppere datasættet i et antal intervaller. Der er ikke nogen facitliste til, hvor mange intervaller der bør være. Det kan være grupperinger à 5 eller à 10, hvis det passer til data. Nogle synes godt om at lave $\sqrt{n}$ intervaller fordelt over variationsbredden, hvor n er antallet af observationer. Denne sidstnævnte metode kan godt blive lidt "kunstig", da man normalt ville forvente grupperinger à 5 eller à 10.

Konklusionen må være, at det er op til det enkelte tilfælde, hvordan det er smartest at inddele variationsbredden, så det giver mest mening.

I dette eksempel grupperes højderne først i intervaller à 10 *cm*. Sidenhen i intervaller à 5 *cm*.

Sættes intervalbredden til 10, bliver variationsbredden delt op i følgende 5 intervaller:
$]150;160],]160;170],]170;180],]180;190]$ og $]190;200]$.

Bemærk intervalgrænserne! F.eks. 160 er kun med i det nederste interval. Det er sådan set ligegyldigt om en værdi er inkluderet i det nedre eller det øvre interval, så længe værdien kun er med én gang. Ellers ville alle med en højde på 160 *cm* jo blive registreret to gange.

Mange større datasæt er kun tilgængelige som grupperede datasæt, og her vil man ofte komme frem til et andet resultat for deskriptorerne, end hvis man havde adgang til kildedata. Her ved man jo ikke, hvordan data er fordelt i de enkelte intervaller, så her vil man principielt altid lave en fejl. Man antager normalt altid at data er fordelt jævnt i et interval. Ser man f.eks. på intervallet 160 til 170, hvor hyppigheden er 10, så vil man antage at hyppigheden er 5 i både intervallet 160 til 165 og fra 165 til 170, men det fremgår jo tydeligt af ovenstående pindediagram, at det bestemt ikke er tilfældet.

Sætning:
Når kildedata ikke er kendt, antages data at være jævnt fordelt i et interval.

Skemaet, som blev brugt i eksemplet med diskrete variable laves igen.

Umiddelbart er den eneste ændring, at der er tilføjet en ekstra kolonne: "Interval midtpunkt: m_i".

Højde i cm $]x_{i-1};x_i]$	Intervalmidtpunkt m_i	Interval hyppighed h_i	Intervalfrekvens f_i	Summeret Frekvens F_i	Produkt $m_i \cdot f_i$
				0,0000	
] 150 ; 160]	155	2	0,0741	0,0741	11,4815
] 160 ; 170]	165	10	0,3704	0,4444	61,1111
] 170 ; 180]	175	10	0,3704	0,8148	64,8148
] 180 ; 190]	185	3	0,1111	0,9259	20,5556
] 190 ; 200]	195	2	0,0741	1,0000	14,4444
		27			172,4

1. kolonne : $]x_{i-1};x_i]$ "Elevernes højde"

De værdier, som undersøgelsen kan antage. I dette tilfælde er det antallet af elevernes højde målt i *cm*. Bemærk igen, at observationerne nu godt kan "falde udenfor ganske bestemte værdier" – f.eks. kan man sagtens være 182 *cm* høj, selvom 182 *cm* ikke er en eksakt observation i skemaet. Så bliver man bare indskrevet i intervallet]180;190].

2. kolonne : $m(x_i)$ "Intervalmidtpunkt"

Her beregnes midtpunkterne af værdi-intervallerne. Dette er nødvendigt for at have en fast værdi at gå ud fra. Det medfører naturligvis også en vis usikkerhed i resultaterne, at man beregner værdierne ud fra intervalmidtpunktet, for man kan jo principielt ikke vide, om alle observationer i et bestemt interval alle ligger over eller under intervalmidtpunktet. I det virkelige liv må man gå ud fra at observationerne er ligeligt fordelt – som allerede nævnt, og derfor kan vi se bort fra denne usikkerhed. Har man dog mistanke om, at datamaterialet er for usikkert, kan man jo indføre mindre intervaller. Dette vil medføre en større præcision i resultaterne.

Intervalmidtpunkterne beregnes ganske enkelt som:

$$m_i = Intervalstartværdi + \frac{Intervalslutværdi - Intervalstartværdi}{2}$$

F.eks. er intervalmidtpunktet for intervallet

$$]160;170] = 160 + \frac{170-160}{2} = 160 + \frac{10}{2} = 160 + 5 = 165$$

Pas på!!! Intervallerne behøver ikke at være lige store. Hverken i opgaverne eller i det virkelige liv. Så lad være med at stole blindt på, at man bare kan tage det samme interval og lægge til det forrige, men i stedet udregne hvert interval med den ovenstående formel.
Her er det vigtigt at vide, at i det tilfælde, hvor intervallerne er uens, så regner man snarere med arealet af "frekvensen", for så kan man operere med forskellige intervaller.

3. kolonne : $h(x_i)$ "Intervalhyppighed"

Hvor mange gange optræder en værdi i observationssættet? Forklaringen er nøjagtig den samme, som for de diskrete variable, dog med den forskel, at hvis der er 20 forekomster af højden 179 *cm* og ikke andre observationer i intervallet]170;180], så vil de 20 forekomster i teorien antages at være jævnt fordelt i hele intervallet.

4. kolonne : $f(x_i)$ "Intervalfrekvens"

Frekvensen eller den relative hyppighed. Forklaringen er nøjagtig den samme, som for de diskrete variable.

5. kolonne : $F(x_i)$ "Summeret Frekvens"

Her tager man frekvenserne og lægger sammen. Igen er forklaringen præcis den samme som for de diskrete variable.

6. kolonne : $m(x_i) \cdot f(x_i)$ "Produkt"

Produktet bruges mest til at finde middelværdi (=gennemsnit). Hver række bidrager til produktet med værdien (i dette tilfælde højden) ganget med frekvensen.
Men da højderne er angivet i intervaller er det nødvendigt at bruge intervalmidtpunkterne og gå ud fra, at observationerne er rimelig ligelig fordelt. Det er dog allerede antaget, så det er ikke noget problem.

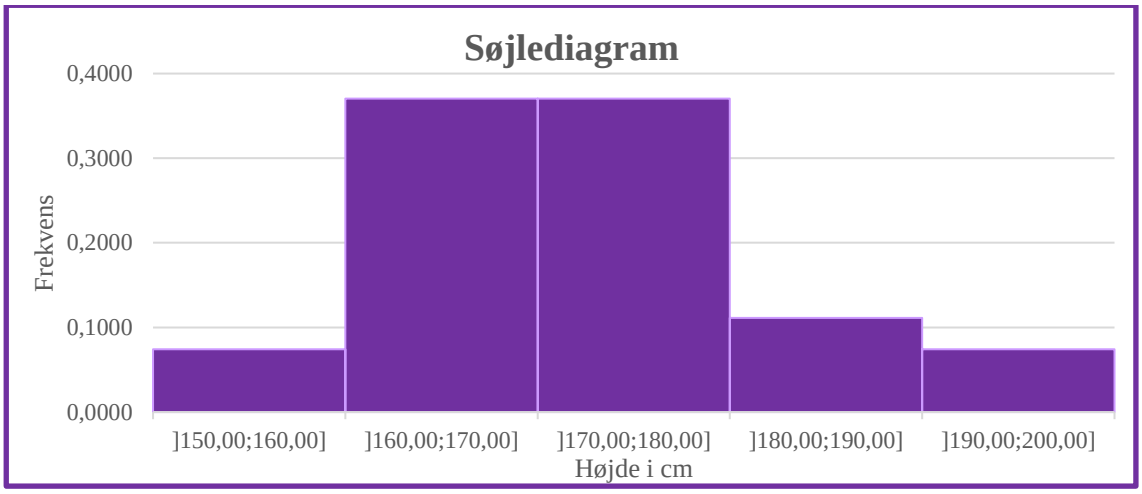
I dette tilfælde ville man sige, at der er 10 elever, som er mellem 160 og 170 *cm* høje. Det vides ikke bedre end at sige at de gennemsnitligt er 165 *cm* høje. Nogle er lavere og nogle er højere, og det går nogenlunde op. De 165 *cm* multipliceres med den udregnede frekvens for de observerede 10 elever. I det samlede regnskab giver dette 11,4845. Når man så har fundet produktet for hver værdi, lægges disse sammen. Det vil altså sige, at man har fundet ud af, at eleverne i hele undersøgelsen i gennemsnit er 172,4 *cm* høje.

Diagrammer for grupperede variable

Typisk vælger man at afbilde data for grupperede variable i to diagrammer. Et **søjlediagram** (kaldes også for et histogram) og en **frekvenskurve**.

Søjlediagrammet laves på samme måde som pindediagrammet. Den eneste forskel er udseendet på søjlerne, idet søjlerne her skal være så brede, at de støder op mod hinanden. Dette gøres for at illustrere at det gælder for HELE intervallet. I søjlediagrammet afsætter man værdierne ud af abscisse-aksen og frekvensen op ad ordinat-aksen.
Bemærk at der – ligesom for de diskrete observationer - gælder for BÅDE søjlediagrammet og frekvenskurven at ordinat-aksen ALTID går mellem 0 og 1. (Frekvenskurven går netop fra 0 til 1, medens søjlediagrammet – som i nedenstående eksempel – kan være endnu mindre, afhængigt af frekvensernes størrelse.)

Deskriptiv statistik

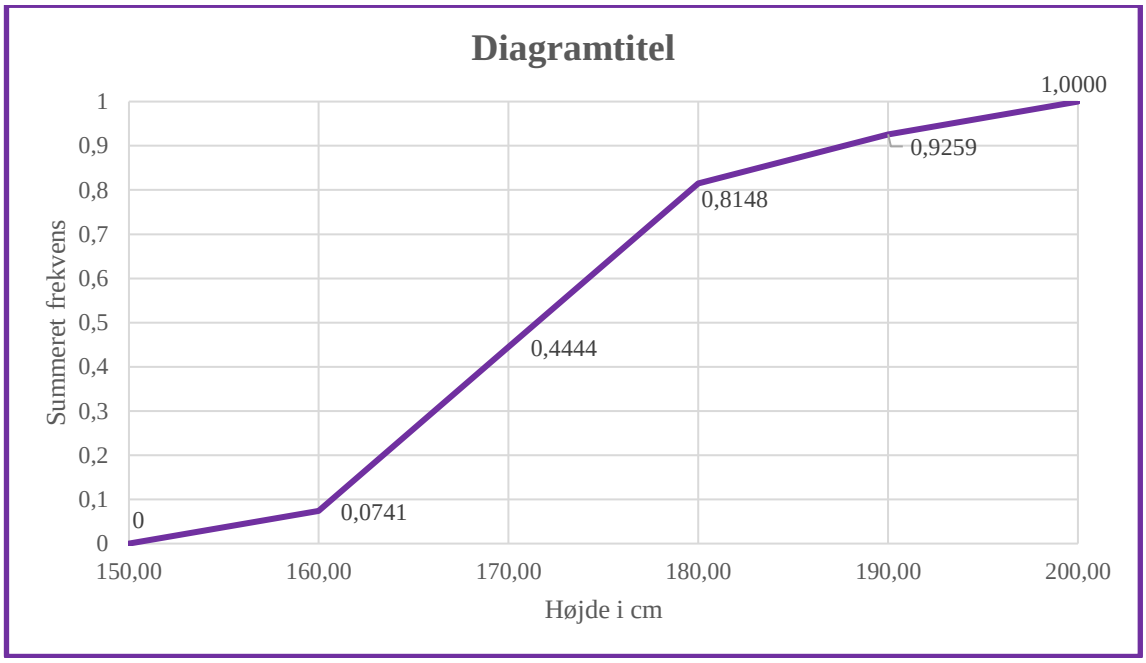


(Her er et diagram, som er lavet i Excel. Bemærk, at i egenskaber for dataserie er mellemrummet mellem søjlerne sat til 0, for at gøre søjlerne så brede som muligt.

Antal observationer: n	27								
Mindste observation:	158								
Største observation:	195								
Intervaller, à:	○ 1	○ 5	◉ 10	○ 50	○ 100	○ 500	○ 1000	○ Kvrod(n)	○ Andet
Hvis "Andet" - Indtast:	200								
Valgt intervalstørrelse:	10,00								
Mindste værdi i tabel:	150								
Største værdi i tabel:	200								
Antal intervaller:	5								

Højde i *cm* $]x_{i-1}\,;\,x_i]$	Interval Nedre grænse	Interval-midtpunkt $m(x_i)$	Interval hyppighed $h(x_i)$	Interval frekvens $f(x_i)$	Summeret frekvens $F(x_i)$	Produkt af hyppighed $m(x_i)\cdot h(x_i)$	Produkt af frekvens $m(x_i)\cdot f(x_i)$
] 150,00 ; 160,00]	150,00	155,00	2	0,0741	0,0741	310	11,48148
] 160,00 ; 170,00]	160,00	165,00	10	0,3704	0,4444	1650	61,11111
] 170,00 ; 180,00]	170,00	175,00	10	0,3704	0,8148	1750	64,81481
] 180,00 ; 190,00]	180,00	185,00	3	0,1111	0,9259	555	20,55556
] 190,00 ; 200,00]	190,00	195,00	2	0,0741	1,0000	390	14,44444

Bemærk, at der i skemaet indsættes en ny kolonne, ”Interval – Nedre grænse”. Den er ikke strengt nødvendig for at løse opgaven, men den kan være en stor hjælp. Specielt hvis frekvenskurven skal indtegnes i Excel, så er den faktisk uundværlig. I frekvenskurven vises den summerede frekvens. Her afsætter man værdierne ud af abscisse-aksen. Sørg for at have den mindste værdi ved ordinat-aksen. Værdien for den summerede frekvens aftegnes som en kurve, som er stykvis lineær (en ret linie) mellem punkterne. Lad kurven begynde i (150,0). Bemærk, at det er hér hvor den ekstra kolonne bruges, således at stregen går fra (150,0) til (160;0,0741) og videre fra (160;0,0741) til (170;0,4444) osv.

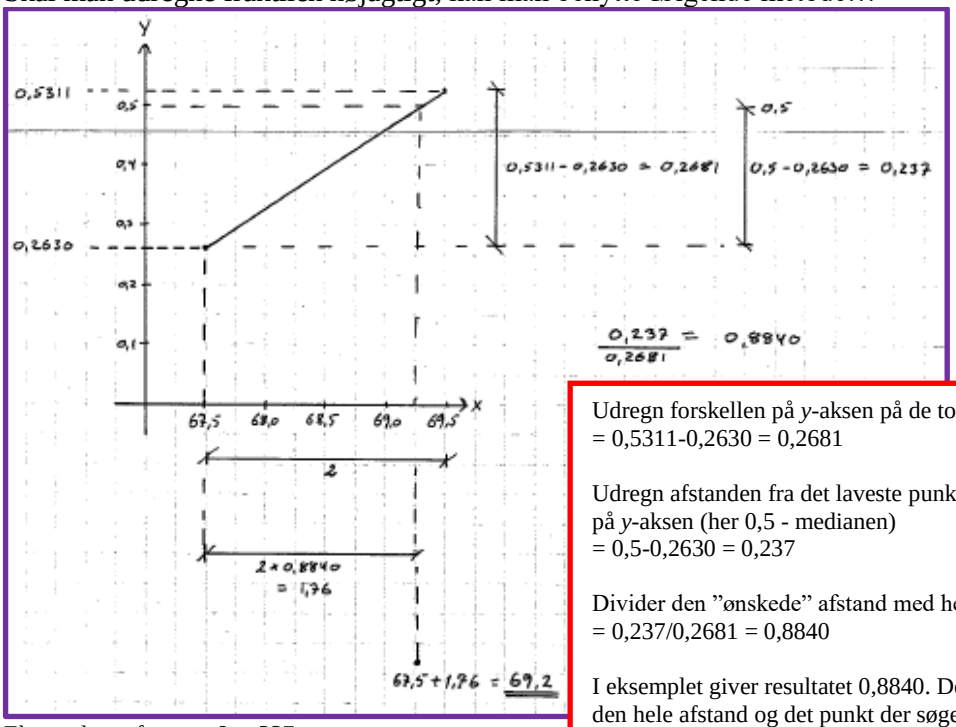


Her er en frekvenskurve, som er lavet i Excel. Bemærk, at punkterne går direkte fra (150,0) og til (160;0,0741) osv. Dette er ikke en Excel standard, for der vil kurven begynde midt mellem 150 og 160. Men hvis man går ind i Excel i diagrammet og højreklikker på *x*-aksen og vælger ”Formater akse”, der kan man vælge at ”Lodret akse krydser” ”På aksemærker”, og så virker det!

Deskriptiv statistik

Mht. fraktiler og dermed også kvartiler, er de lidt sværere at aflæse – både i tabellen og kurven. Har man indtegnet kurven på millimeterpapir, kan man lave en ok aflæsning.

Skal man udregne fraktilen nøjagtigt, kan man benytte følgende metode…



Eksemplet er fra opg. 8 p. 337

Udregn forskellen på *y*-aksen på de to endepunkter.
= 0,5311-0,2630 = 0,2681

Udregn afstanden fra det laveste punkt til det ønskede punkt på *y*-aksen (her 0,5 - medianen)
= 0,5-0,2630 = 0,237

Divider den "ønskede" afstand med hele afstanden.
= 0,237/0,2681 = 0,8840

I eksemplet giver resultatet 0,8840. Dette er forholdet mellem den hele afstand og det punkt der søges.

Nu kendes forholdet lodret. Det er det samme forhold vandret, så afstanden væk fra den mindste værdi beregnes:

$= 2 \cdot 0{,}8840 = 1{,}76 \quad (2 \text{ er den vandrette afstand } = 69{,}5 - 67{,}5)$

Men der tages jo udgangspunkt i $x = 67{,}5$, så den ønskede værdi for medianen er: $= 67{,}5 + 1{,}76 = 69{,}2$

Denne sidste øvelse vil dog ofte være overflødig.
Dette er en af de få matematiske discipliner, hvor det kan være mere formålstjenstligt at aflæse end at beregne, eller få det udregnet i CAS.

Husk, at data i forvejen er approksimerede, idet data antages at være jævnt fordelt i hvert interval. Derfor må det påpeges at den ekstra præcision der er givet ved at beregne de eksakte fraktiler ikke kan ændre på den usikkerhed, som er iboende i måden at behandle data på.

Derfor vil det være ekstremt få tilfælde, hvor denne metode vil komme i brug.