3. Statistiske parametre.

3.1. Innledning.

I det forrige notatet så vi på teknikker for å illustrere viktige egenskaper ved et datasett grafisk. Men vi trenger også noen tall som raskt kan gi oss informasjon om slike egenskaper, og som kan sette oss i stand til å sammenlikne datasett. I dette notatet skal vi se på to grupper av slike statistiske parametre: mål for midtpunkt og mål for spredning.

Det kan være nyttig å huske at i de fleste tilfellene er vårt datasett et *utvalg* fra en større *populasjon*. Selv om vi foretar våre beregninger på data fra utvalget, vil vi som regel bruke parametrene til å si noe om den populasjonen som utvalget er hentet fra. Våre parametre blir *estimater* for de tilsvarende parametrene i populasjonen.

3.2. Mål for midtpunkt.

En av de viktigste størrelsene som karakteriserer et datasett, er *midtpunktet* i datasettet. Dette kan defineres på flere forskjellige måter. De vanligste er ***middelverdi*** (også kalt ***aritmetisk gjennomsnitt*** eller bare ***gjennomsnitt***) og ***median***, men det fins også andre mål for midtpunkt.

3.2.1. Middelverdi.

Vi definerer:

Middelverdien $\bar{x}$ for et sett av n dataelementer $x_1, x_2, \ldots, x_n$ er definert som

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Med andre ord: summer alle dataene og del på antall dataelementer. Det er vanlig å angi middelverdien med en desimal mer enn i de opprinnelige dataene.

Eksempel 3.1: Finn middelverdien for de 6 tallene 2, 4, 7, 10, 15, 21.

Løsning: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{6}(2+4+7+10+15+21) = \frac{1}{6}\cdot 59 \approx \underline{\underline{9.8}}$.

På tilsvarende måte definerer vi middelverdien i en *populasjon*. Vi bruker den greske bokstaven μ for å symbolisere *populasjonens* middelverdi. Dersom populasjonen består av N elementer, er

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \,.$$

3.2.2. Median.

Vi definerer:

Medianen for et sett av *n* dataelementer er verdien til det midterste elementet i det *sorterte* datasettet. Vi bruker gjerne $\tilde{x}$ som symbol for median.

Dersom *n* er et oddetall, går det greit å finne medianen etter denne oppskriften. Medianen blir da verdien til element nr $\frac{n+1}{2}$. Men dersom *n* er et jamt tall, bruker vi *gjennomsnittet* av tall nr $\frac{n}{2}$ og tall nr $\frac{n+1}{2}$.

Eksempel 3.2: Finn medianen for:
a) De 5 tallene 4, 15, 7, 2, 10.
b) De 6 tallene 7, 21, 10, 2, 15, 4.

Løsning: Vi starter med å sortere tallene i stigende rekkefølge:
a) Medianen for tallene 2, 4, 7, 10, 15 er tall nr $\frac{5+1}{2}=3$, slik at $\tilde{x}=7$.
b) Medianen for tallene 2, 4, 7, 10, 15, 21 er middelverdien av tall nr 3 og nr 4, slik at

$$\tilde{x}=\frac{7+10}{2}=8.5.$$

3.2.3. Andre mål for midtpunkt.

Selv om middelverdi og median er de mest brukte målene for midtpunkt, finnes det også andre slike mål. Noen eksempler:

- ***Midrange*** er midtpunktet mellom høyeste og laveste verdi i det sorterte datasettet.
- ***Modus*** er den verdien som forekommer oftest. Brukes helst dersom dataene kun kan ha noen få bestemte verdier. Hvis for eksempel en avis vurderer ukas filmer og gir dem "terningkast" fra 1 til 6, vil modus-verdien være den verdien som forekommer oftest. Denne verdien er da et mål for kvaliteten på ukas filmer.

3.2.4. Sammenfatning.

Dersom dataene er noenlunde "normalt" fordelt uten at noen av dataene har "ekstreme" verdier, er *middelverdien* det beste målet for midtpunkt. Men dersom det fins data som har "ekstreme" verdier (såkalte "outliers"), kan det være bedre å bruke medianen. Hvis du for eksempel skal angi gjennomsnitts formue for en gruppe personer, og en av dem er skipsreder med milliardformue mens de andre er vanlige lønnsmottakere, vil medianen være et bedre mål for "typisk" formue enn middelverdien.

Vi kan vise at middelverdien for data fra et *utvalg* er det beste estimatet for middelverdien for den *populasjonen* som utvalget er hentet fra. Noe tilsvarende gjelder ikke for de andre målene vi har for midtpunkt. Dette gjør at under normale forhold er middelverdien det nyttigste målet for midtpunkt.

3.3. Mål for spredning.

Det er også viktig å ha mål for *spredningen* av dataverdiene i et datasett. Jeg skal først se på *standardavviket* som er det klart mest brukte spredningsmålet. Deretter skal jeg se på noen andre spredningsmål som er mindre brukt, selv om de er enklere å finne enn standardavviket.

3.3.1. Varians og standardavvik.

Før vi går løs på standardavviket, skal vi definere begrepet ***varians***. Vi må da skille mellom variansen for en *populasjon* og variansen for et *utvalg* fra populasjonen:

Dersom en *populasjon* består av *N* data med verdier $x_1,\ x_2 .\ \ldots,\ x_N$, er ***populasjonsvariansen*** gitt ved

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

der μ er middelverdien i populasjonen.

Dersom et *utvalg* består av *n* data med verdier $x_1,\ x_2 .\ \ldots,\ x_n$, er ***utvalgsvariansen*** gitt ved

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

der $\bar{x}$ er middelverdien i utvalget.

La oss se nærmere på hva disse formlene inneholder. Uttrykket $x_i - \mu$ eller $x_i - \bar{x}$ angir hvor mye en dataverdi avviker fra middelverdien. For alle dataverdiene skal dette avviket beregnes. Deretter skal alle disse avvikene kvadreres, og kvadratene skal summeres. Til slutt deles på *N* eller på $n-1$.

Hvorfor har vi to nesten like formler? Grunnen er at vi kan vise at dersom variansen for et utvalg beregnes etter den gitte formelen, får vi det beste estimatet av variansen i den populasjonen som utvalget er hentet fra. Og siden vi vanligvis må nøye oss med utvalg, er formelen for utvalgsvariansen den klart mest brukte. Dersom det kun er tale om "varians" uten nærmere presisering, er det så å si alltid formelen for utvalgsvarians som brukes.

Formlene for varians er noe kronglete å bruke fordi vi først må beregne alle avvikene fra middelverdien, deretter kvadrere alle disse avvikene, og til slutt summere. Vi slipper lettere unna dersom vi lager varianter av formelen. Nå får vi bruk for denne kvadratsummen til andre formål senere, så vi kan like godt døpe den S_{XX} (merk *stor S*) med en gang og lage uttrykk som er mer hensiktsmessig i bruk:

$$S_{XX} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \underline{\sum_{i=1}^{n} x_i^{\,2} - n\cdot(\bar{x})^2} = \underline{\sum_{i=1}^{n} x_i^{\,2} - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$$

Merk at:

$\sum_{i=1}^{n} x_i^{\,2}$ betyr at du *først* kvadrerer tallene, og *deretter* summerer kvadratene.

$\left(\sum_{i=1}^{n} x_i\right)^2$ betyr at du *først* summerer tallene, og *deretter* kvadrerer summen.

I begge de to siste formene av formelen slipper vi å beregne avvikene $x_i - \bar{x}$. Vi nøyer oss med å kvadrere de opprinnelige dataene og summere disse kvadratene.

Den midterste formen ser penest ut, men jeg vil anbefale at du bruker den siste formen. Grunnen er at dersom $\bar{x}$ er unøyaktig angitt (for eksempel ved avrunding), vil $n\cdot(\bar{x})^2$ bli enda mer unøyaktig. Og siden det ofte viser seg at $n\cdot(\bar{x})^2$ er nesten like stor som $\sum_{i=1}^{n} x_i^{\,2}$, kan selv en liten unøyaktighet i ett av leddene føre til en stor unøyaktighet i differensen S_{XX}.

Når vi har funnet S_{XX}, finner vi lett variansen som

$$s^2 = \frac{1}{n-1}S_{XX} \qquad \text{eller} \qquad \sigma^2 = \frac{1}{N}S_{XX}\,.$$

Eksempel 3.3: Bruk de tre versjonene av varians-formelen til å beregne variansen for tallene 8, 10 og 13.

Løsning: Vi finner først middelverdien med en desimal:

$\bar{x} = \frac{1}{3}(8+10+13) = \frac{1}{3}\cdot 31 = \underline{10.3}$.

Da blir

$$s^2 = \frac{1}{n-1}S_{XX} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{3-1}\left((8-10.3)^2+(10-10.3)^2+(13-10.3)^2\right)$$
$$= \tfrac{1}{2}(5.29+0.09+7.29) = \underline{\underline{6.335}}$$

$$s^2 = \frac{1}{n-1}S_{XX} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^{\,2} - n\cdot(\bar{x})^2\right) = \frac{1}{3-1}\left((8^2+10^2+13^2)-3\cdot 10.3^2\right)$$
$$= \tfrac{1}{2}(64+100+169-3\cdot 106.09) = \underline{\underline{7.365}}$$

$$s^2 = \frac{1}{n-1}S_{XX} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^{\,2} - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right) = \frac{1}{3-1}\left((8^2+10^2+13^2)-\frac{1}{3}\cdot 31^2\right)$$
$$= \tfrac{1}{2}\left(64+100+169-\tfrac{1}{3}\cdot 961\right) = \underline{\underline{6.333}}$$

Vi ser at den første og den siste versjonen av formelen gir nesten samme resultat, mens den midterste versjonen gir et avvikende resultat. Påvis selv at den siste versjonen er minst følsom for avrundings-unøyaktigheter!

Utledning av de tre variantene av formelen:

$$\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2 \overset{(1)}{=} \sum_{i=1}^{n}\left(x_i^{\,2}-2x_i\cdot\overline{x}+\left(\overline{x}\right)^2\right) \overset{(2)}{=} \sum_{i=1}^{n}x_i^{\,2}-2\overline{x}\cdot\sum_{i=1}^{n}x_i+\underbrace{\left(\overline{x}\right)^2+\left(\overline{x}\right)^2+\cdots+\left(\overline{x}\right)^2}_{n\text{ ganger}}$$

$$\overset{(3)}{=}\sum_{i=1}^{n}x_i^{\,2}-2\overline{x}\cdot\left(n\cdot\overline{x}\right)+n\cdot\left(\overline{x}\right)^2 \overset{(4)}{=} \underline{\sum_{i=1}^{n}x_i^{\,2}-n\cdot\left(\overline{x}\right)^2} \overset{(5)}{=} \underline{\sum_{i=1}^{n}x_i^{\,2}-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2}$$

Kommentarer:

(1): Kvadrerer ut hvert ledd med 2. kvadratsetning.

(2): Stokker om leddene slik at den felles faktoren $\overline{x}$ kan settes utenfor parentes (d.v.s. utenfor summetegnet).

(3): Benytter at $\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i \quad\Leftrightarrow\quad \sum_{i=1}^{n}x_i=n\cdot\overline{x}$.

(4): Ser at $-2\overline{x}\cdot\left(n\cdot\overline{x}\right)=-2n\cdot\left(\overline{x}\right)^2$, slik at de to siste leddene kan trekkes sammen til $-\left(\overline{x}\right)^2\cdot n$.

(5): Setter inn at $\left(\overline{x}\right)^2=\frac{1}{n^2}\left(\sum_{i=1}^{n}x_i\right)^2$ og forkorter bort n.

Etter at vi har funnet variansen (som regel finner vi utvalgs-variansen), beregner vi ***standardavviket*** ved å trekke kvadratrota av variansen:

Standardavviket er

$\sigma=\sqrt{\sigma^2}$ for en populasjon, $s=\sqrt{s^2}$ for et utvalg

Dermed får standardavviket samme benevning som middelverdien.

Eksempel 3.4: Beregn standardavviket for de tre tallene 8, 10 og 13.

Løsning: Vi vet fra forrige eksempel at variansen er $s^2=6.33$. Da blir standardavviket

$$s=\sqrt{s^2}=\sqrt{6.33}\approx\underline{\underline{2.52}}.$$

3.3.2. Kvartiler og prosentiler.

Vi definerer *nedre* og *øvre kvartil* etter samme prinsipp som vi brukte da vi definerte median:

Nedre kvartil er den dataverdien som er slik at 25 % av de sorterte dataelementene har verdi mindre enn eller lik nedre kvartil.

Øvre kvartil er den dataverdien som er slik at 75 % av de sorterte dataelementene har verdi mindre enn eller lik øvre kvartil.

Disse definisjonene virker tilsynelatende greie – helt til du prøver å bruke dem i praksis. Da viser det seg ofte at det ikke er noe veldefinert dataelement som tilfredsstiller kravet for et kvartil. Vi må på en eller annen måte interpolere mellom dataverdier på samme måte som vi gjorde da vi fant medianen for et jamt antall elementer. Det er dessverre ingen allment anerkjent metode for å gjøre dette, og lærebøker (og dataprogram) har ulike metoder som kan gi ulike resultater. Vi skal derfor ikke gå nærmere inn på disse detaljene.

Når vi har funnet de to kvartilene, definerer vi *interkvartilbredden* slik:

Interkvartilbredden er avstanden mellom øvre kvartil og nedre kvartil.

Dersom vi ønsker andre prosenttall enn 25 %, kan vi bruke ***prosentiler***. De defineres (fremdeles noe unøyaktig) slik:

n % av dataene i datasettet har en verdi som er lavere enn eller lik ***n*-prosentilet**.

Kvartilene blir da henholdsvis 25- og 75-prosentilene, mens medianen blir 50-prosentilet.

Når vi skal bestemme median, kvartiler og prosentiler, starter vi alltid med å sortere dataene slik eksemplet nedenfor viser.

Eksempel 3.5: Bestem median, nedre og øvre kvartil, og interkvartilbredden for eksamens-poengene fra notatet om grafiske framstillinger:

44	51	24	24	44	55	44	42	41	11	100	83
58	41	42	37	56	24	20	49	20	40	43	31
41	26	64	35	15	9	37	45	7	24	17	41
6	9	52	42	10	53	20	43	83	54	16	83
35	66	19	44	50	22	30	57	65	63	48	36
22	79	49	25	65	50	35	63	21	43	55	24
44	55	17	14	62	28	33	8	67	55	32	42
37	8										

Løsning: Når vi sorterer dataene, får vi:

6	7	8	8	9	9	10	11	14	15	16	17
17	19	20	20	20	21	22	22	24	24	24	24
24	25	26	28	30	31	32	33	35	35	35	36
37	37	37	40	41	41	41	41	42	42	42	42
43	43	43	44	44	44	44	44	45	48	49	49
50	50	51	52	53	54	55	55	55	55	56	57
58	62	63	63	64	65	65	66	67	79	83	83
83	100										

Vi har 86 dataelementer. I det sorterte materialet har vi rammet inn element nr. 43 og 44, som gir at medianen blir

$$\tilde{x} = \tfrac{1}{2}(41+41) = \underline{\underline{41}}.$$

For å finne kvartilene, benytter vi at

$$0.25 \cdot 86 = 21.5.$$

Vi rammer derfor også inn element nr 21 og 22, og element nr 65 og 66. Da får vi:

Nedre kvartil: $\frac{1}{2}(24+24) = \underline{\underline{24}}$.

Øvre kvartil: $\frac{1}{2}(53+54) = \underline{\underline{53.5}}$.

Interkvartilbredden blir

$53.5 - 24 = \underline{\underline{29.5}}$.

For moro skyld fikk jeg Excel til å beregne median, 25prosentil (nedre kvartil) og 75prosentil (øvre kvartil) for disse dataene. Resultatene ble henholdsvis 41, 24 og 52.75. De to første verdiene var som ventet, mens verdien for øvre kvartil viser at definisjonene av kvartiler er noe flytende.

3.3.3. Andre spredningsmål.

De spredningsmålene vi har sett på hittil, er de mest brukte. Spesielt standardavviket er mye brukt. Men standardavviket krever mye regnearbeid. Dersom du kun er ute etter et raskt mål for spredningen, er *rangen* grei å ty til:

Rangen er avstanden mellom høyeste og laveste dataverdi.

I vårt eksempel blir rangen

$100 - 6 = \underline{\underline{94}}$.

Midlere absolutt avvik fra middelverdien er kanskje et mer naturlig mål for spredning. Vi finner da absoluttverdien av hver dataverdis avstand fra middelverdien, og beregner deretter middelverdien av alle disse avstandene. Med formel:

Midlere absolutt avvik fra middelverdien er

$$\frac{1}{n}\sum_{i=1}^{n}\left|x_i - \overline{x}\right|.$$

Denne formelen er imidlertid klønete i bruk på grunn av absoluttverditegnene, noe som gjør at den brukes lite. Vi *må* ha med absoluttverditegnene, ellers vil summen gi null som resultat fordi

$$\sum_{i=1}^{n}\left(x_i - \overline{x}\right) = \sum_{i=1}^{n} x_i - \left(\underbrace{\overline{x} + \overline{x} + \cdots + \overline{x}}_{n \text{ ganger}}\right) = n\cdot\overline{x} - n\cdot\overline{x} = \underline{0}$$

der vi har benyttet at

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \Leftrightarrow \quad \sum_{i=1}^{n} x_i = n\cdot\overline{x}\,.$$

3.4. Klasseinndelt materiale.

3.4.1. Innledning.

Hittil har vi forutsatt at vi har alle dataene tilgjengelige når vi beregner parametrene. Men dersom vi kun har tabeller med klasseinndelt materiale tilgjengelig, blir beregningene litt annerledes. Problemet er at vi mister detalj-informasjon fordi vi ikke lenger vet den eksakte verdien av hvert enkelt dataelement. Vi vet bare hvor mange dataelementer som faller i hver klasse, samt klassegrensene. Dette fører til at beregningene blir mindre nøyaktige. Heldigvis fører det også til at beregningene blir enklere.

3.4.2. Beregning av middelverdi og varians.

Når vi beregner *middelverdi* og *varians* for et klasseinndelt materiale, antar vi at alle data-elementene i en klasse har samme verdi, nemlig *klassemidtpunktet*. For kontinuerlige data definere dette som midtpunktet mellom øvre (nedre) klassegrense og øvre (nedre) klasse-grense i naboklassen. For diskrete data er det best å definere klassemidtpunktet som midt-punktet mellom klasseskillene.

Vi forutsetter altså at alle dataene i en klasse har samme verdi, nemlig klassemidtpunktet. Istedenfor å summere disse dataverdiene, multipliserer vi antall dataelementer i hver klasse med klassemidtpunktet. Vi kaller midtpunktet i klasse nr. j for x_j mens antall elementer i klassen er f_j. Antall dataelementer er n, mens antall klasser er K. Da blir:

Gruppert middelverdi: $\bar{x}_g = \frac{1}{n}\sum_{j=1}^{K} f_j x_j$.

Vi får en nyttig variant av denne formelen dersom vi plasserer $\frac{1}{n}$ innenfor summetegnet:

$$\bar{x}_g = \frac{1}{n}\sum_{j=1}^{K} f_j x_j = \sum_{j=1}^{K} \frac{f_j}{n} x_j \, .$$

Brøken $\frac{f_j}{n}$ er den *relative frekvensen* for klasse nr j, eller andelen av alle dataelementene i denne klassen.

Vi beregner *variansen* for et gruppert materiale på tilsvarende måte:

Gruppert varians:

$$\begin{aligned} s^2{}_g &= \frac{1}{n-1}\sum_{j=1}^{K} f_j \left(x_j - \bar{x}_g\right)^2 \\ &= \frac{1}{n-1}\left(\sum_{j=1}^{K} f_j {x_j}^2 - n\cdot\left(\bar{x}_g\right)^2\right) \\ &= \frac{1}{n-1}\left(\sum_{j=1}^{K} f_j {x_j}^2 - \frac{1}{n}\cdot\left(\sum_{j=1}^{K} f_j x_j\right)^2\right) \end{aligned}$$

Jeg har bare tatt med formlene for utvalgs-varians. Du får formlene for populasjons-varians ved å erstatte nevneren $n-1$ med *N*.

Etter at du har funnet variansen, finner du standardavviket ved å trekke kvadratrota.

Som eksempel skal vi se på de poengsummene til eksamen som vi har benyttet tidligere. Disse dataene ligger i et regneark, og jeg benytter regnearket både til å beregne eksakte verdier av middelverdi og varians, og til å beregne gruppert middelverdi og gruppert varians.

Poeng	Midtpunkt x_j	Frekvens f_j	$f_j \cdot x_j$	$f_j \cdot x_j^2$
1 - 10	5,5	7	39	211,75
11 - 20	15,5	10	155	2402,50
21 - 30	25,5	12	306	7803,00
31 - 40	35,5	11	391	13862,75
41 - 50	45,5	22	1001	45545,50
51 - 60	55,5	11	611	33882,75
61 - 70	65,5	8	524	34322,00
71 - 80	75,5	1	76	5700,25
81 - 90	85,5	3	257	21930,75
91 - 100	95,5	1	96	9120,25
Summer:		86	3453	174782

Nå får vi:

Gruppert middelverdi:

$$\overline{x}_g = \frac{1}{n}\sum_{j=1}^{K} f_j x_j = \frac{1}{86}\cdot 3453 = \underline{\underline{40.15}}.$$

Gruppert varians:

$$s^2{}_g = \frac{1}{n-1}\left(\sum_{j=1}^{K} f_j x_j{}^2 - \frac{1}{n}\cdot\left(\sum_{j=1}^{K} f_j x_j\right)^2\right) = \frac{1}{86-1}\left(174782 - \frac{1}{86}\cdot 3453^2\right) = \underline{\underline{425.18}}.$$

Til sammenlikning kan vi beregne middelverdi og varians for det opprinnelige (ugrupperte) materialet. Da får vi middelverdi $\overline{x} = 39.81$ og varians $s^2 = 402.56$. Vi ser at vi får litt unøyaktige resultater ved å bruke grupperte data. Det er naturlig siden vi antar at alle dataelementene i en klasse har samme verdi (nemlig klassemidtpunktet) istedenfor å bruke de eksakte verdiene.

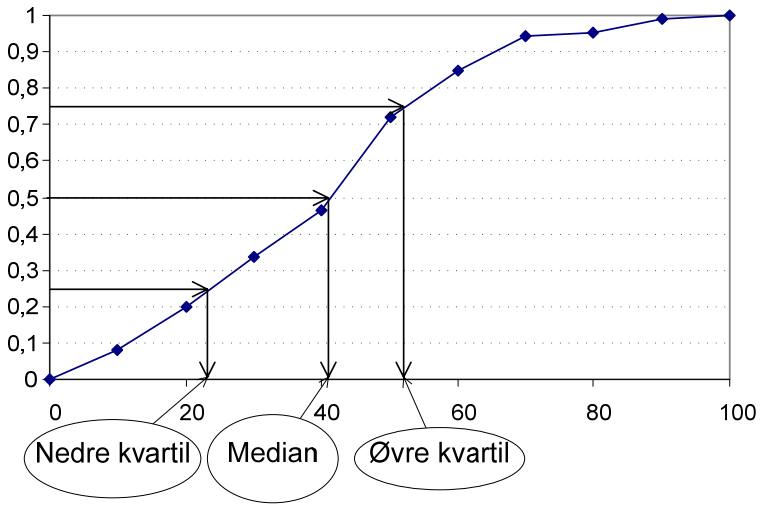
Dette illustrerer at det ikke har noen hensikt å operere med en mengde desimaler. En tommelfinger-regel er at under mellomregninger skal du bruke minst to sifre mer enn antall sifre i rådataene, unntatt når du beregner kvadratsummene. Da må du regne så nøyaktig som mulig. Resultatene angis vanligvis med ett siffer mer enn i rådataene.

3.4.3. Median, kvartiler og prosentiler for klasseinndelt materiale.

Da vi regnet ut middelverdi og varians for klasseinndelt materiale, antok vi at alle dataelementene i en klasse har samme verdi (klassemidtpunktet). Men når vi regner ut *median*, *kvartiler* og *prosentiler*, antar vi at alle dataverdiene i en klasse er *jevnt fordelt*

innenfor sin klasse. Da kan vi bruke grafene til de *kumulerte relative frekvensene* til å finne disse størrelsene grafisk.

Framgangsmåten er vist nedenfor med utgangspunkt i den grafen for kumulert relativ frekvens for eksamens-poengene som vi har kommet fram til tidligere:



Du finner medianen ved å trekke ei horisontal linje fra 50%-punktet på aksen for relativ kumulert frekvens (*y*-aksen) og bort til grafen, og deretter ei linje ned til førsteaksen. Der denne linja treffer førsteaksen er medianen. På vår figur blir medianen litt over 40, la oss si ca. 41. På tilsvarende måte finner du at nedre kvartil er litt over 20 (la oss si ca. 23) og øvre kvartil er litt over 50 (la oss si ca. 52). Da blir interkvartilbredden omtrent

$$52 - 23 = 29.$$

Disse verdiene er ikke langt fra de eksakte verdiene som vi har funnet tidligere på grunnlag av enkeltdataene.

Dersom du skal finne prosentiler generelt, går du fram på samme måte.

Det er fullt mulig å regne mer eksakt ved at du først bruker grafen til å finne ut hvilket intervall du befinner deg i. Deretter setter du opp likningen for den rette linja i dette intervallet på grunnlag av kjente start- og sluttpunkter. Da kan du finne den *x*-verdien som svarer til en ønsket *y*-verdi. Men det er sjelden nødvendig å foreta slike beregninger, og vi skal ikke benytte slike metoder nå.

3.5. Sammenfatning.

Standardavviket er det mest brukte målet for hvor stor spredningen er i et datasett. Vi kan vise at dersom dataene er noenlunde "normalt" fordelt innenfor datasettet, har vi at:

- Om lag to tredeler av alle dataene i et datasett ligger mindre enn ett standardavvik fra middelverdien.
- Om lag 95 % av alle dataene i et datasett ligger mindre enn to standardavvik fra middelverdien.

Disse påstandene skal vi komme nærmere tilbake til senere i kurset.

La oss sjekke om dette stemmer med dataene i vårt eksempel. Vi har tidligere funnet at middelverdien er $\bar{x} = 39.81 \approx 40$, mens standardavviket er $s = \sqrt{s^2} = \sqrt{402.56} \approx 20$. Da skal om lag $\frac{2}{3}$ av dataene ligge innenfor

$$[\bar{x} - s, \bar{x} + s] = [40 - 20, 40 + 20] = [20, 60].$$

Av grafen for kumulert relativ frekvens ser vi at ca. 20% av dataene ligger under $x = 20$, mens ca. 85% av dataene ligger under $x = 60$. Da vil ca $85\% - 20\% = 65\%$ ligge mellom disse verdiene, og det er jo ganske nær $\frac{2}{3}$.

Et dataelements plassering i utvalget eller i populasjonen angis gjerne med elementets ***standard score*** eller ***z-score***, som defineres slik:

Standard score (***z-score***) er:

$$z_i = \frac{x_i - \bar{x}}{s} \text{ for et dataelement i et utvalg.}$$

$$z_i = \frac{x_i - \mu}{\sigma} \text{ for et dataelement i en populasjon.}$$

z-scoren angir *hvor mange standardavvik* en dataverdi er fra middelverdien. Positiv *z*-score betyr at dataverdien er større enn middelverdien, mens negativ *z*-score betyr at dataverdien er mindre enn middelverdien. Ut fra det vi nettopp har sagt om fordelingen av dataverdier, kan vi slå fast at for et "normalt" datasett har om lag to tredeler av dataene *z*-score mellom -1 og 1, mens om lag 95% av dataene har *z*-score mellom -2 og 2.

Dette betyr at dersom et dataelement har $z < -2$ eller $z > 2$, så har dette elementet en "unormal" verdi. I vårt eksempel har minste og største dataverdi *z*-scorer på henholdsvis

$$z_{\min} = \frac{6 - 40}{20} = \underline{-1.7},$$

$$z_{\max} = \frac{100 - 40}{20} = \underline{3.0}.$$

Vi ser at den største dataverdien ($x = 100$) ligger påfallende langt fra middelverdien.

Dataelementer der $|z| > 3$ bør betraktes med skepsis. Det kan hende at et slikt dataelement har sneket seg inn i vårt datasett ved en feil.

Hittil har vi sett hvordan et enkelt dataelement plasserer seg i datasettet. Nå er tiden inne til å se nærmere på datasettet som helhet. Da er det viktig å merke seg at standardavviket må ses i sammenheng med middelverdien. Dersom standardavviket er lite sammenliknet med middelverdien, har vi et datasett med liten spredning. Men er standardavviket stort sammenliknet med middelverdien, har vi stor spredning. Slike vurderinger har stor praktisk interesse. Dersom du produserer bolter, og boltenes diameter varierer litt med et standardavvik på 0.1 mm, kan dette være helt uakseptabelt for tynne bolter med diameter på 0.5 cm. Men et slikt standardavvik kan være helt OK for tykke bolter med diameter på 10 cm.

Et utvalg eller en populasjon kan karakteriseres ved sin ***variasjonskoeffisient*** (”Coefficient of Variation”, CV), som defineres slik:

Variasjonskoeffisienten er

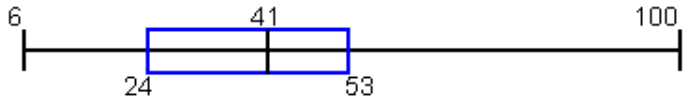
$CV = \frac{s}{\overline{x}}$ for et utvalg, $CV = \frac{\sigma}{\mu}$ for en populasjon.

CV kan oppfattes som *relativ spredning rundt middelverdien*. En liten verdi for *CV* betyr at spredningen er liten sammenliknet med middelverdien. Hvis for eksempel et datasett har en *CV* på 0.01, betyr det at standardavviket er 1 % av middelverdien, noe som igjen betyr at ca. to tredeler av alle dataene ligger mindre en 1 % fra middelverdien.

For vårt Excel-eksempel blir

$$CV = \frac{s}{\overline{x}} = \frac{20}{40} = \underline{0.50}\,.$$

Et ***boxplot*** er en figur som gir et visuelt inntrykk av hvordan dataene i et datasett er fordelt. Et boxplot for dataene i vårt eksempel er vist nedenfor:



En linje med veldefinerte endepunkter angir omfanget av dataverdiene. På denne linja er det en boks som avgrenses av nedre og øvre kvartil. Gjennom boksen er medianen tegnet inn. Dersom det er noen ”ekstreme” verdier, blir også disse spesielt angitt. Den ”ekstreme” verdien på 100 i vårt datasett faller sammen med endepunktet og blir derfor ikke spesielt framhevet på denne figuren.