Sentralmål
og
spredningsmål

av

Peer Andersen

Sentralmål og spredningsmål i statistikk

I dette notatet skal vi se på de viktigste momentene om sentralmål og spredningsmål slik de blir brukt i statistikken.

Sentralmål

Vi har flere typer sentralmål. De mest vanlige er gjennomsnitt, median og typetall. Alle disse er pensum i grunnskolen. Her er noen eksempler på hvordan vi beregner disse målene. Vi tar utgangspunkt i følgende tall.

2, 3, 4, 1, 0, 3, 2, 1, 4, 5, 3

Gjennomsnittet beregnes ved å legge sammen alle verdiene og deretter dele på antall observasjoner. Gjennomsnittet betegnes ofte med $\bar{x}$.

$$\bar{x} = \frac{2+3+4+1+0+3+2+1+4+5+3}{11} = 2{,}54$$

Medianen finner vi å ordne tallene i stigende rekkefølge og deretter ta den midterste verdien. Dersom det er et partall med observasjoner tar vi gjennomsnittet av de to i midten.

0, 1, 1, 2, 2, 3, 3, 3, 4, 4, 5

Medianen er i dette tilfelle 3.

Typetallet er den verdien som inntreffer flest ganger. I dette tilfellet ser vi at det er tallet 3. Dersom vi har to verdier som opptrer like ofte varierer det litt hva litteraturen sier. Noen bøker sier at det ikke finnes typetall, mens andre sier at det da er to (eller flere) typetall.

Fordeler og ulemper med de ulike sentralmålene

Vi skal her se på noen fordeler og ulemper med de ulike sentralmålene. For å kunne beregne gjennomsnitt og median må vi ha tallstørrelser. Det er ikke nødvendig for typetallet. La oss si vi er på fisketur og får 3 torsk, 4 sei og 5 lyr. Her gir det ingen mening å snakke om gjennomsnitt eller median. Vi kan derimot si at lyr er det vi har fått mest av. (Typetall)

Gjennomsnittet lar seg påvirke av ekstreme verdier. Det gjør ikke medianen eller typetallet. La oss se på følgende eksempel som viser lønnsnivået i en liten bedrift. Her er det de ansatte tjener:

320 000, 280 000, 350 000, 310 000, 1 500 000

De 4 laveste verdiene er lønnen til arbeiderne, mens lønnen på 1 500 000 kroner er sjefen sin lønn

Den gjennomsnittlige inntekten i bedriften er

$$\bar{x} = \frac{320000+280000+350000+310000+1500000}{5} = 552\,000$$

Medianen derimot er 320 000. Hva synes du gir den beste beskrivelsen av lønnsnivået i bedriften? Som vi ser påvirkes gjennomsnittet av ekstreme verdier. Det gjør ikke medianen i samme grad, og den gir nok et bedre bilde av lønnsnivået enn gjennomsnittet. Typetallet påvirkes heller ikke av ekstreme verdier. En ulempe med typetallet er som i eksempelet over. Her er alle observasjonene forskjellige og det gir ikke noe mening å snakke om typetallet

Spredningsmål

Som vi så i sted så har vi flere ulike sentralmål for å beskrive et datamateriale. Disse målene gir oss verdifull informasjon om materialet, men som vi snart skal se så er ikke dette nok for å gi en god beskrivelse. La oss se på følgende eksempel. I et fengsel i Norge ble det en gang arrangert en sommerskole slik at de innsatte skulle ha et tilbud om meningsfylt aktivitet også i sommermånedene. I etterkant ble det gjort en undersøkelse om hva de innsatte syntes om sommerskolen. Vi kan anta at det var 10 innsatte som svarte på dette. De skulle gi poeng der 1 var dårligst og 5 var best. Gjennomsnittsverdien ble 3 og vi kan anta at alle 10 svarte. Hva kan vi si om datamaterialet? Her er det flere muligheter. La oss se på tre ulike datasett som alle gir et gjennomsnitt på 3.

Alt. 1 3, 3, 3, 3, 3, 3, 3, 3, 3, 3

Alt. 2 1, 5, 1, 5, 1, 5, 1, 5, 1, 5

Alt. 3 1, 3, 2, 4, 5, 4, 5, 2, 1, 3

Alle disse tre seriene gir et gjennomsnitt på 3. Likevel representerer de et nokså forskjellig materiale. Det er stor forskjell på hvordan vi skal følge opp saken om alle svarer at de er middels fornøyd som i alternativ 1 i forhold det som alternativ 2 viser der de enten er svært misfornøyd eller svært godt fornøyd med tilbudet. Alternativ 3 er noe mitt mellom med noen som er godt fornøyd, noen som er passe fornøyd og noen som mindre fornøyd. Det som er åpenbart er at gjennomsnittet alene ikke er nok for å gi en god beskrivelse av et datamateriale. Vi kommer tilbake til hva vi kan gjøre litt senere. La oss se på et eksempel til.

Gjennomsnittsalderen på en sydendestinasjon er 22 år. Ville du reist til denne plassen om du selv var rundt 22 år og var ute etter en festetur med jevnaldrede? Ville du reist til denne plassen om du var mellom 30 og 40 år og hadde et par unger? Dette spørsmålet er faktisk ikke mulig å gi noe godt svar på kun ut i fra gjennomsnittet. La oss se på to destinasjoner der snittalderen i begge tilfeller er 22 år.

Destinasjon A 20, 24, 23, 21, 18, 26, 24, 20

Destinasjon B 4, 40, 6, 38, 10, 34, 3, 41

På begge disse stedene er snittalderen 22 år. Skulle jeg reist til syden med mine to 7 åringer ville jeg åpenbart valgt destinasjon B. En som er ute etter fest og moro sammen med andre yngre personer ville nok valgt destinasjon A. Igjen ser vi at gjennomsnittet ikke er nok til å gi en god beskrivelse av materialet.

For å kunne gi en bedre beskrivelse av et datamateriale er det vanlig å også oppgi et spredningsmål. Spredningsmålet er et mål for hvor stor spredningen er i materialet. Det finnes flere ulike spredningsmål. Vi skal her se på noen av dem.

Variasjonsbredde

Variasjonsbredde er et enkelt spredningsmål og også det eneste spredningsmålet som er med i grunnskolens pensum. Det går ganske enkelt ut på et en ser på differansen mellom den største og minste verdien. På destinasjon A ser vi at variasjonsbredden er 26 – 18 = 8. På destinasjon B ser vi at variasjonsbredden er 41 – 3 = 38. Ved å oppgi variasjonsbredden ser vi at på destinasjon A er observasjonene ganske konsentrerte mens det er stor spredning på destinasjon B. Fordelen med variasjonsbredde er at det er enkelt i bruk og enkelt å regne ut. Ulempen er at det ofte ikke gir noe godt bilde av situasjonen. Tenk deg at på destinasjon A er det en av personene på f. eks 24 år som har med ungen sin på 1 år. Det ville gitt et stort utslag på variasjonsbredden selv om plassen fremdeles ville vært dominert av ungdommer i 20 årene.

Kvartil, kvartildifferanse og kvartilavvik

Disse begrepene er definert noe ulikt fra lærebok til lærebok. En ser f. eks at QED definerer det annerledes enn det Alfa gjør. Knut Ole Lysø bruker i sine bøker samme definisjon som QED. Dette kan selvsagt være forvirrende. Vi skal se på begge disse metodene. Til eksamen er det selvsagt nok å kunne en av definisjonene.

<u>Alfa sin definisjon av kvartiler</u>

Slik Alfa definerer dette så blir 2. kvartil definert som medianen på samme måte som vi definerte medianen innledningsvis. Medianen deler datasettet i to deler. Tar vi den nederste delen og regner ut medianen til den nederste delen finner vi første kvartil. Tilsvarende finner vi tredje kvartil ved å regne ut medianen til øverste delen av datasettet. La oss se på et eksempel for å belyse dette. Vi tar utgangspunkt i et eksempel med 11 data. Dette kan være alder på en gruppe mennesker. Vi sorter disse i stigende rekkefølge

5, 12, 17, 19, 23, 24, 26, 34, 43, 46, 50

Medianen er det midterste tallet som vi ser er 24 i dette tilfelle. Det delen av datasettet som er mindre enn medianen er

5, 12, 17, 19, 23

For å finne 1. kvartil tar vi medianen til disse dataene. Som vi ser er den 17 og 1. kvartil er med andre ord 17. På tilsvarende måte finner vi 3. kvartil ved å regne ut medianen til dataene

26, 34, 43, 46, 50

Den er som vi ser 43 og 3. kvartil er med andre ord 43.

Vi ser til slutt på eksempelet med to de sydendestinasjonene. Her har vi sortert dataene i stigende rekkefølge for begge destinasjonene

Destinasjon A 18, 20, 20, 21, 23, 24, 24, 26

Destinasjon B 3, 4, 6, 10, 34, 38, 40, 41

Vi startet med destinasjon A. Vi ser at medianen her er gjennomsnittet av fjerde og femte data som er 22. Det nedre delen av datasettet består av dataene

18, 20, 20, 21

Vi ser her at medianen og dermed 1. kvartil er 20. På tilsvarende måte finner vi 3. kvartil ved å beregne medianen til dataene

23, 24, 24, 26

Som vi ser så er den 24.

På tilsvarende måte kan vi regne ut median og kvartilene til destinasjon B. Vi ser medianen er 22. Videre kan vi regne ut 1. kvartil til å være 5 og 3. kvartil til å være 39.

QED og Lysø sin definisjon av kvartiler

QED og Lysø har en litt annen måte å definere kvartiler på. Første kvartil er definert som den verdien som er slik at en firedel av observasjonene er mindre eller lik denne verdien. Andre kvartil er definert slik at halvparten av verdiene er mindre eller lik denne verdien. Andre kvartil er i praksis det samme som medianen. Tredje kvartil definerer vi som den verdien som er slik at tre firedeler av verdiene er mindre eller lik denne observasjonen.

La oss se på hva kvartilene blir på de to sydendestinasjonene. Først ordner vi de i stigende rekkefølge.

Destinasjon A 18, 20, 20, 21, 23, 24, 24, 26

Destinasjon B 3, 4, 6, 10, 34, 38, 40, 41

På destinasjon A ser vi at første kvartil er 20, andre kvartil er 21 og tredje kvartil er 24.
På destinasjon B ser vi at første kvartil er 4, andre kvartil er 10 og tredje kvartil er 38.

I dette eksempelet var det greit å finne kvartilene siden vi tilfeldigvis hadde 8 observasjoner. Hva gjør vi hvis vi har et antall observasjoner som ikke er delelig på 4? La oss si at vi har 57 observasjoner. Tar vi 25 % av dette får vi ikke noe heltall men derimot 14,25. Vi bruker vanlige avrundingsregler og finner at 1. kvartil er den 14. observasjonen. Tilsvarende finner vi at 2. kvartil er 28,5 som vi runder av til 29. Andre kvartil er med andre ord den 29. observasjonen. Regner vi ut hva 3. kvartil får vi 42,75. Dette runder vi opp til 43 slik at tredje kvartil blir observasjon nummer 43.

Kvartildifferanse og kvartilavvik

Differansen mellom tredje og første kvartil er det vi kaller kvartilbredden. Kvartilavviket kan defineres som halvparten av kvartilbredden. På sydendestinasjonene vil kvartilbredden på A være 4 og kvartilavviket vil være 2. På destinasjon B vil kvartilbredden og kvartilavviket være henholdsvis 34 og 17. Vi ser at kvartilbredden og kvartilavviket i dette eksempelet blir det samme om vi legger til grunn Alfa sin definisjon og om vi bruke QED/Lysø sin definisjon. Lysø bruker i sine bøker kvartildifferanse istedenfor kvartilbredde. Dere skal ellers være oppmerksomme på at det finnes flere ulike definisjoner av kvartilavvik.

Vi kan tolke kvartilbredden til å være spredningen på de 50 % observasjonene som ligger i midten. Eller sagt på en annen måte, så er kvartilbredden det samme som variasjonsbredden når vi tar bort de 25 % minste og 25 % største observasjonene.

Gjennomsnittlig absoluttavvik

Et annet mål er det vi kaller gjennomsnittlig absoluttavvik. Det går ut på at vi regner ut hvor stort avviket er fra gjennomsnittet for alle observasjonene. Deretter regner vi ut gjennomsnittet av alle avvikene. La oss se på eksempelet med sydendestinasjonene.

Destinasjon A

$$GA_A = \frac{|20-22| + |24-22| + |23-22| + |21-22| + |18-22| + |26-22| + |24-22| + |20-22|}{8} =$$

$$GA_A = \frac{2+2+1+1+4+4+2+2}{8} = 2{,}25$$

(Når vi setter dette tegnet | foran og etter en differanse betyr det at vi skal ta forskjellen mellom tallene. Det vil se det største tallet minus det minste tallet.)

Destinasjon B

$$GA_B = \frac{|4-22| + |40-22| + |6-22| + |38-22| + |10-22| + |34-22| + |3-22| + |41-22|}{8} =$$

$$GA_B = \frac{18+18+16+16+12+12+19+19}{8} = 16{,}25$$

Vi ser her at på destinasjon A er hver person i gjennomsnitt enten 2,25 år yngre eller eldre enn snittalderen på 22 år. På destinasjon B ser vi at hver person i gjennomsnitt er 16,25 år eldre eller yngre enn snittalderen på 22 år. Ved å oppgi gjennomsnittet og det gjennomsnittlige absoluttavviket får vi en god beskrivelse av hvordan datamaterialet ser ut. Gjennomsnittlig absoluttavvik er et godt mål for spredning men det er likevel lite brukt i praksis. I praktisk matematikk er det et mål som heter standardavviket som er enerådende. Det skal vi se på nå.

Standardavvik

Standardavviket er som sagt det spredningsmålet som brukes mest i praktisk matematikk. Det skyldes at det har en rekke interessante egenskaper som er nyttig i videre arbeidet med statistikken. Dette ligger imidlertid utenfor vårt pensum og vi skal her nøye oss med å se på hvordan vi kan regne ut standardavviket og hvordan vi kan tolke det.

Standardavviket beregnes ved at vi først regner ut avviket mellom hver enkelt observasjon og gjennomsnittet. Disse kvadreres og deretter summerer vi de sammen. Vi deler så på antall observasjoner og til sist tar vi kvadratroten. La oss se på dette gjennom eksempelet med sydendestinasjonene. Standardavviket betegnes gjerne med s. Vi ser først på destinasjon A

$$s_A = \sqrt{\frac{(20-22)^2 + (24-22)^2 + \cdots + (24-22)^2 + (20-22)^2}{8}} =$$

$$s_A = \sqrt{\frac{2^2+2^2+1^2+1^2+4^2+4^2+2^2+2^2}{8}} = \sqrt{\frac{4+4+1+1+16+16+4+4}{8}} =$$

$$s_A = \sqrt{\frac{50}{8}} = 2{,}5$$

Vi ser nå på destinasjon B.

$$s_B = \sqrt{\frac{(4-22)^2 + (40-22)^2 + \cdots + (3-22)^2 + (41-22)^2}{8}} =$$

$$s_B = \sqrt{\frac{18^2+18^2+16^2+16^2+12^2+12^2+19^2+19^2}{8}} =$$

$$s_B = \sqrt{\frac{324+324+256+256+144+144+361+361}{8}} =$$

$$s_B = \sqrt{\frac{2170}{8}} = 16{,}47$$

Vi ser at det kan være ganske omstendelig å regne ut standardavviket. Er det mange observasjoner må vi i praksis bruke dataverktøy for å få dette til.

<u>Tolkning av standardavvik</u>

Det store spørsmålet som nå melder seg er hvordan vi skal tolke standardavviket. Dette er ikke så helt enkelt, men jeg skal prøve å komme med noen betraktninger om dette. Det første vi legger merke til er at forskjellene mellom standardavvik og gjennomsnittlig absoluttavvik er forholdsvis små. I praksis betyr det at vi ikke gjør noe stor feil om vi sier at standardavviket er tilnærmet det samme

som det gjennomsnittlige absoluttavviket. Det vil si hvor mye observasjonene i snitt avviker fra gjennomsnittet.

En annen interessant egenskap ved standardavviket har vi når vi har et normalfordelt materiale. Da vil 68,2 % av observasjonene ligge innenfor gjennomsnittet minus ett standardavvik og gjennomsnittet pluss et standardavvik. Tilsvarende kan en vise at 95,4 % av observasjonene ligger innenfor gjennomsnittet minus to standardavvik og gjennomsnittet pluss to standardavvik. Det ligger langt utenfor pensum i matematikk 1 og gå inn på en begrunnelse for dette. La oss imidlertid se på et lite eksempel for å belyse dette.

La oss si at høyden på alle menn i Norge i gjennomsnitt er 180 cm og at standardavviket er 10 cm. Høyden til menn kan vi anta er normalfordelt. Da vil 68,2 % av mennene i Norge være mellom 170 og 190 cm. Tilsvarende kan vi også si at 95,4 % av mennene vil være mellom 160 cm og 200 cm.

Når en får presentert statistiske undersøkelser er det ofte vanlig å bare oppgi gjennomsnitt, standardavvik og antall observasjoner og så blir det vår jobb og tolke disse. La oss se på et eksempel. Vi tar for oss to skoleklasser hver på 30 elever. I klasse A er gjennomsnittscoren på en matematikkprøve 45 poeng av 70 mulige. Standardavviket er 5 poeng. I klasse B er også gjennomsnittsscoren 45 poeng men her er standardavviket 16,5 poeng. Hva kan vi si om nivået i matematikk i disse to klassene når vi legger denne prøven til grunn? Jo vi ser at gjennomsnittet er det samme men standardavviket er det ganske stor forskjell på. I klasse A er det mye mindre enn i klasse B. Det betyr at elevene er mye jevnere i A klassen enn i B klassen. Det er mange som scorer rundt gjennomsnittet i A klassen og det er få som har svart riktig bra og det er også få som har svart meget svakt. I B klassen er derimot standardavviket mye større og det er mye større spredning i klassen. Her vil en finne flere riktig gode besvarelser og også en del tilsvarende dårlige besvarelser og relativt få som scorer rundt gjennomsnittet. I tabellene under er det gitt et eksempel på poeng som gir de beskrevne verdiene.

A klassen

35	38	39	40	40	40	41	41	42	42
43	43	44	44	45	45	45	46	46	46
46	47	48	48	49	50	51	54	55	57

B klassen

10	15	16	21	24	25	26	27	32	40
45	46	46	49	50	55	55	56	57	57
57	57	58	58	59	60	60	60	64	65

Det å kunne beskrive et datasett ut i fra gjennomsnitt og standardavvik er viktig å kunne. Det forventes selvsagt ikke at en skal gi eksakte beskrivelser av datasettene, men at en kan si noe om hvordan dataene er fordelt slik jeg har gjort i teksten over tabellene.

La oss ta et eksempel til. Etter en konsert ble publikum spurt om hva de syntes om konserten. De kunne gi poeng fra 1 til 6 der 1 var svært dårlig og 6 var meget bra. Gjennomsnittet var 3,5 og standardavviket var 2. Hva kan vi si om publikums opplevelse av konserten. Vi ser at snittet er

akkurat mitt mellom 1 og 6. Vi legger også merke til at vi har et standardavvik på 2. Det er høyt når skalaen går fra 1 til 6. Her vil vi kunne trekke slutningen at de fleste som har svart har enten gitt en høy score eller en lav score og at det er få som har krysset av for middels. Sagt på godt norsk så har flertallet av deltakerne på konserten enten ment at den var meget bra eller så har de ment at den var dårlig. Få av deltakerne mente at den var middels. Hadde gjennomsnittet vært 3,5 og standardavviket vært 0,2 kunne vi derimot trukket slutningen at her var det store flertallet sånn passe fornøyd med konserten og at få var meget fornøyd og også at få syntes den var veldig dårlig.