

Deskriptiv statistik - ugrupperede observationer

Varians og spredning

Fra: "Vejen til matematik, AB1+C"

I matematikken har man følgende formelle definitioner:

Definition 1.1: Middelværdi

Ved middelværdien \bar{x} for et datasæt $x_1, x_2, x_3, \dots, x_n$ forstås man:

$$\begin{aligned}\bar{x} &= \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} \\ &= \frac{1}{n} \sum_{i=1}^n x_i\end{aligned}$$

Her har vi i sidste linje anvendt det såkaldte *summationstegn* Σ (det græske bogstav sigma), der ofte bruges til at angive, at tal skal lægges sammen. Bogstavet i er en variabel, som her løber fra 1 til n . Man får det første led i summen ved at sætte $i = 1$, andet led ved at sætte $i = 2$ osv. Det sidste led fremkommer for $i = n$.

Oftentimes har man brug for at kunne udregne, hvordan de observerede værdier spredt sig omkring middelværdien. Til dette formål benytter man *varians* og *standardafvigelse* (spredning):

Definition 1.2: Varians og standardafvigelse

Ved *variansen* for et datasæt $x_1, x_2, x_3, \dots, x_n$ forstås man middelværdien af de kvadrerede afvigelser fra middelværdien:

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

Ved *standardafvigelsen* for et datasæt forstås man kvadratroden af variansen:

$$\sigma(x) = \sqrt{Var(x)}$$

1. Ikke-grupperede observationer

Eksempel 1.3: Middelværdi og standardafvigelse

Søren har fået karaktererne 00, 4, 7, 10.
Vi beregner middelværdien (gennemsnittet), variansen og standardafvigelsen:

$$\bar{x} = \frac{0+4+7+10}{4} = 5,25$$

$$\text{Var}(x) = \frac{1}{4} \sum (x_i - \bar{x})^2$$

$$= \frac{(0-5,25)^2 + (4-5,25)^2 + (7-5,25)^2 + (10-5,25)^2}{4} = 13,69$$

$$\sigma(x) = \sqrt{\text{Var}(x)} = \sqrt{13,69} = 3,70$$

Fordi flere af Søren's karakterer ligger langt fra middelværdien, bliver standardafvigelsen stor, næsten 4 karakterer.

Øvelse 1.4: Flere karakterer

Beregn middelværdi, varians og standardafvigelse for dine egne terminkarakterer.

Ved skriftlig eksamen i matematik blev der i 2009 givet over 9000 karakterer. I undervisningsministeriet og rundt omkring på skolerne er man meget interesseret i at få et overblik over resultatet, der foreligger i form af denne tabel:

Karakter	x	-3	00	02	4	7	10	12
Antal elever	hyp.	81	1094	579	1727	2912	2034	615

I stedet for en tabel med over 9000 tal har man valgt at nøjes med at angive, hvor mange gange hver karakter forekommer. Tallene i første række (karaktererne) kaldes *observationer*, og dem i anden række (antal) kaldes den observerede *hyppighed*. Som så ofte i matematikken, kaldes en vilkårlig karakter for x . Karakter nr. i kaldes x_i og dens hyppighed kaldes h_i .

Observationssættets størrelse er 9042, fordi der er givet 9042 karakterer og typekarakteren er 7, fordi det er den karakter, der har den største hyppighed.