

Middelværdi, spredning og skævhed

Et datasæt på generaliseret form kan skrives ved:

X = (x_1, x_2, ..., x_n) hvor x_1, x_2, ..., x_n \in \mathbb{X}

Middelværdien \mu af datasættet beregnes ud fra gennemsnittet ved:

\mu = \frac{x_1 + x_2 + ... + x_n}{n}

Det bemærkes, at der summeres over alle elementerne i datasættet og gennemsnittet kan derfor kun beregnes hvis man kender hele datasættet. Gennemsnittet kan dog generaliseres til begrebet om en middelværdi.

Hvis man i stedet for selve datasættet kender fordelingen af datasættet, dvs. dataværdierne for observationsmængden \mathbb{X} = \{x_1, x_2, ..., x_m\} samt den tilhørende observationsfrekvens f_1, f_2, ..., f_m, kan middelværdien \mu beregnes ved brug af et vægtet gennemsnit, hvor der summeres over samtlige observationsværdier vægtet med den tilhørende observationsfrekvens:

\mu = x_1 \cdot f_1 + x_2 \cdot f_2 + ... + x_m \cdot f_m

Det bemærkes her, at såfremt alle dataværdier behandles som observationsværdier vil antallet af observationværdier være lig med størrelsen af datasættet m = n. Hyppigheden af hver dataværdi vil i så fald blive 1, hvilket vil resultere i, at samtlige dataværdier vil have frekvensen af \frac{1}{n}. Formlen for middelværdien vil i så fald omskrives til formelen for gennemsnittet. For en grupperet fordeling hvor frekvensen f_i er angivet kan middelværdien beregnes ved at anvende midterværdien i intervallet for x_i.

Middelværdien repræsenterer det 1. statistiske moment. Begrebet moment er helt generelt en matematisk metode til at transformere en fordeling til et sæt af værdier, der beskriver fordelingsens egenskaber ud fra en slags ligevægtsprincip. I dette perspektiv angiver det 1. moment det midtpunkt hvor differencen mellem alle dataværdier og midtpunktet summerer til nul. Af denne grund anvendes det 1. moment ligeledes til at bestemme massemidtpunktet (også kaldet tyngdepunktet) af et fysisk objekt hvilket er en god analogi til at skabe sig en intuitiv forståelse af hvad middelværdien repræsenterer.

Vars og spredning

Det 2. statistiske moment er den såkaldte vars som ofte betegnes ved \sigma^2 hvor \sigma kaldes for datasættets spredning. Som navnene antyder angiver variansen (og spredningen) hvor meget et datasæt varierer omkring dets middelværdi.

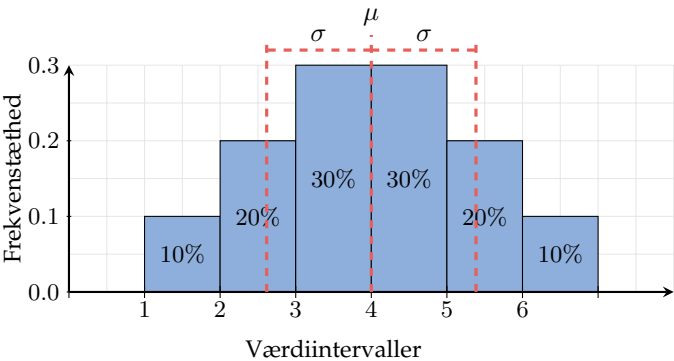
Variansen kan beregnes ud fra datasættet ved:

\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + ... + (x_n - \mu)^2}{n}

Hvis man i stedet for selve datasættet kender fordelingen af datasættet, dvs. dataværdierne for observationsmængden \mathbb{X} = \{x_1, x_2, ..., x_m\} samt den tilhørende observationsfrekvens f_1, f_2, ..., f_m, kan man beregne variansen ved:

\sigma^2 = (x_1 - \mu)^2 \cdot f_1 + (x_2 - \mu)^2 \cdot f_2 + ... + (x_m - \mu)^2 \cdot f_m

Ud fra variansen bestemmes spredningen ved \sigma = \sqrt{\sigma^2}. Spredningen \sigma har samme enhed som selve dataværdierne og kan derfor tolkes som et mål for den afstand som dataværdierne i gennemsnit afviger fra middelværdien. Spredningen er derfor også typisk anvendt når man vil kvantificere måleusikkerheden for en målemetode.



Skævhed

Den tredje egenskab som man typisk anvender til at beskrive en fordeling er hvorvidt fordelingen er symmetrisk eller asymmetrisk omkring middelværdien. Denne egenskab kan undersøges ved at beregne fordelingsens skævhed.

Hvis datasættets median og middelværdi ikke er samme værdi kaldes datasættet skævt. Datasættets skævhed \gamma kan kvantificeres til en talværdi, der kan bestemmes ved:

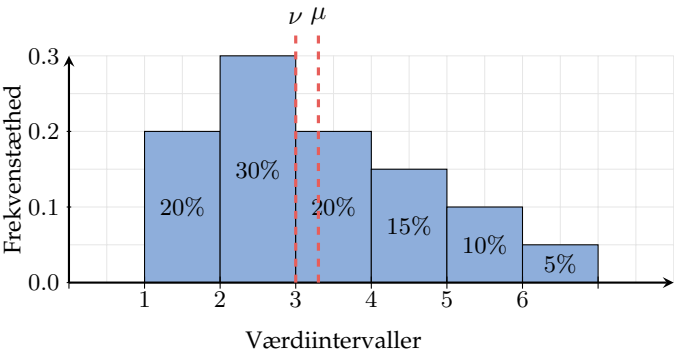
\gamma = \frac{\mu - \nu}{\sigma}

Her er \mu datasættets middelværdi, \nu er datasættets medianværdi, og \sigma er datasættets spredning.

Skævheden er en talværdi i intervallet -1 < \gamma < 1. Hvis datasættets middelværdi er større end medianen er skævheden positiv og fordelingen kaldes højreskæv. Tilsvarende hvis datasættets middelværdi er mindre end medianen er skævheden negativ og fordelingen kaldes venstreskæv.

Skævhed kan således sammenlignes med en vippe hvor vippens tippepunkt placeres i medianen. Hvis middelværdien er større end medianen vil vippen altså tippe til højre da datasættets tyngdepunkt ligger på denne side af tippepunktet. På tilsvarende vis vil vippen tippe til venstre når middelværdien er mindre end medianen.

I figuren nedenfor ses en fordeling med en positiv skævhed hvor middelværdien er større end medianværdien.



Geogebra: mean(...) Variance(...)

Først indtastes data i Geogebbras regneark (spreadsheet).

Hvis man har selve datasættet indtastet i felterne A1 til A100 kan man beregne middelværdien og variansen ved følgende kommandoer i Algebra-vinduet:

```
mean(A1 : A100)
Variance(A1 : A100)
```

Hvis man i stedet har en fordelingstabel med dataværdierne i felterne A1 til A10 og de tilhørende frekvenser i felterne B1 til B10 anvendes kommandoerne:

```
mean(A1 : A10, B1 : B10)
Variance(A1 : A10, B1 : B10)
```