

Middelværdi, spredning og skævhed

Den mest anvendte statistiske deskriptor er *middelværdien* som i daglig tale ofte kaldes *gennemsnittet*. Hvis man f.eks. betragter følgende datasæt:

X = (1, 1, 2, 3, 3, 3, 4, 4, 4, 5)

Idet alle datasættets værdier er kendte kan middelværdien μ af datasættet direkte beregnes ved ud fra gennemsnittet ved:

μ = (1 + 1 + 2 + 3 + 3 + 3 + 4 + 4 + 4 + 5) / 10 = 30 / 10 = 3

Beregningen af middelværdien ud fra datasættet kan generaliseres ved brug af en generaliseret notation for et datasæt med n dataværdier:

X = (x1, x2, ..., xn)

Den tilsvarende generaliserede formel for middelværdien er således:

μ = (x1 + x2 + ... + xn) / n

Det bemærkes, at der summeres over alle elementerne i datasættet og middelværdien kan derfor kun beregnes direkte ved gennemsnittet hvis man kender hele datasættet.

Gennemsnitsberegningen kan dog omskrives til et såkaldt *vægtet gennemsnit* som kan anvendes hvis man kun har kendskab til datasættets hyppighedsfordeling. For ovenstående datasæt er den tilsvarende hyppighedsfordeling:

Dataværdi	1	2	3	4	5
Hyppighed	2	1	3	3	1

Summen af alle dataværdierne kan beregnes ved at gange hyppighederne med dataværdierne for alle m dataværdier og lægge disse sammen, hvilket giver følgende beregning:

μ = (2 · 1 + 1 · 2 + 3 · 3 + 3 · 4 + 1 · 5) / 10 = (2 + 2 + 9 + 12 + 5) / 10 = 30 / 10 = 3

Den tilsvarende generaliserede formel for middelværdien beregnet ud fra hyppighederne $h_1, h_2, ..., h_m$ er således:

μ = (h1 · x1 + h2 · x2 + ... + hm · xm) / n

Bemærk her, at der anvendes n kan beregnes som summen af hyppighederne $n = h_1 + h_2 + ... + h_m$.

Middelværdien kan også beregnes hvis man i stedet for hyppigheden har kendskab til datasættets frekvensfordeling. For at se dette omskrives det vægtede gennemsnit af hyppighederne ved:

μ = (h1 · x1 + h2 · x2 + ... + hm · xm) / n = (h1 · x1 / n + h2 · x2 / n + ... + hm · xm / n) = (h1 / n · x1 + h2 / n · x2 + ... + hm / n · xm) = f1 · x1 + f2 · x2 + ... + fm · xm

I sidste omskrivning anvendes, at frekvensen beregnes ved $f_i = \frac{h_i}{n}$.

For ovenstående datasæt er den tilsvarende frekvensfordeling:

Dataværdi	1	2	3	4	5
Frekvens	0.2	0.1	0.3	0.3	0.1

Middelværdien beregnes nu ved et vægtet gennemsnit hvor hver dataværdi er vægtet med den tilhørende frekvens:

μ = 0.2 · 1 + 0.1 · 2 + 0.3 · 3 + 0.3 · 4 + 0.1 · 5 = 3

For en generel formel anvendes dataværdierne $x_1, x_2, ..., x_m$ samt den tilhørende observationsfrekvens $f_1, f_2, ..., f_m$, hvorved middelværdien μ beregnes ved:

μ = f1 · x1 + f2 · x2 + ... + fm · xm

Det bemærkes her, at hvis 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 formlen for gennemsnittet.

For en grupperet fordeling hvor dataværdierne er grupperet i intervaller kan middelværdien beregnes ved at anvende midterværdien i intervallet som x_i der vægtes med den tilsvarende frekvens f_i for gruppen. Til eksempel betragtes følgende grupperede fordeling hvor midterværdien for værdiintervallet også er angivet:

Værdiinterval	[0, 1[	[1, 2[	[2, 3[	[3, 4[	[4, 5[
Midterværdi	0.5	1.5	2.5	3.5	4.5
Frekvens	0.2	0.1	0.3	0.3	0.1

Middelværdien i dette tilfælde beregnes ud fra intervallernes midterværdi, der vægtes med den tilsvarende frekvens for gruppen:

μ = 0.2 · 0.5 + 0.1 · 1.5 + 0.3 · 2.5 + 0.3 · 3.5 + 0.1 · 4.5 = 2.5

Vars og spredning

Det 2. statistiske moment er den såkaldte *varians* som ofte betegnes ved σ^2 hvor σ 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:

σ² = ((x1 - μ)² + (x2 - μ)² + ... + (xn - μ)²) / n

Hvis man i stedet for selve datasættet kender hyppighedsfordelingen af datasættet, dvs. dataværdierne $\mathbb{X} = \{x_1, x_2, ..., x_m\}$ samt den tilhørende hyppighed $h_1, h_2, ..., h_m$, kan man beregne variansen ved:

σ² = (h1 · (x1 - μ)² + h2 · (x2 - μ)² + ... + hm · (xm - μ)²) / n

Hvis man i stedet for hyppigheden kender dataværdiernes frekvenser $f_1, f_2, ..., f_m$, kan man beregne variansen ved:

σ² = (x1 - μ)² · f1 + (x2 - μ)² · f2 + ... + (xm - μ)² · fm

Ud fra variansen bestemmes spredningen ved $\sigma = \sqrt{\sigma^2}$. Spredningen σ 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.

Middelværdi, vars og spredning

Et datasæt er givet med dataværdierne $x_1, x_2, ..., x_m$ og de tilhørende frekvenser $f_1, f_2, ..., f_m$.

Datasættets middelværdi μ beregnes heraf ved:

μ = f1 · x1 + f2 · x2 + ... + fm · xm

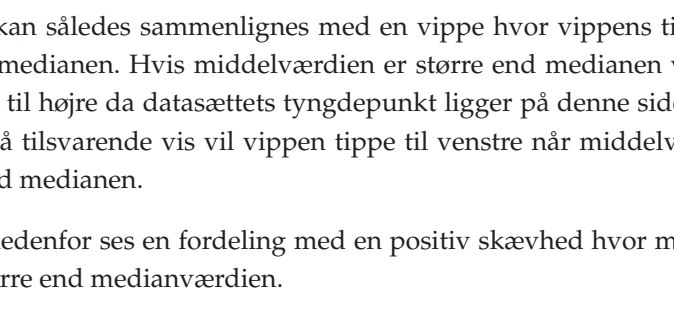
Datasættets vars σ^2 beregnes ved:

σ² = (x1 - μ)² · f1 + (x2 - μ)² · f2 + ... + (xm - μ)² · fm

Spredningen σ beregnes ud fra variansen ved:

σ = √σ²

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 fordelings 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.



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 fordelings *skævhed*.

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

γ = (μ - ν) / σ

Her er μ datasættets middelværdi, ν er datasættets medianværdi, og σ 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)