

Spredning

Spredning

Vi genopfrisker definitionen for middelværdi fra sidst.

Definition 1.1 (Middelværdi). Middelværdien for en stokastisk variabel X med værdimængde $\{x_1, x_2, \dots, x_n\}$ er defineret ved

$$\mu = x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) \cdots + x_n \cdot P(X = x_n).$$

Middelværdien giver os et indblik i, hvad den stokastiske variabel mest sandsynligt antager af værdi, men den siger ikke noget om, hvor meget værdierne, variabelen kan antage varierer omkring middelværdien μ . Vi vil definere variansen som den ”gennemsnitlige variation fra middelværdien”. Derfor ville det give mening at definere variansen som

$$\text{Var} = \mathbb{E}(X - \mu),$$

men dette vil tage højde for, at $X - \mu$ nogle gange er negativ og nogle gange er positiv. Vi vil bare gerne vide, hvor langt X ligger fra μ , så derfor definerer vi variansen på følgende vis:

Definition 1.2 (Varians og spredning). Lad X være en (diskret) stokastisk variabel med værdimængde $\{x_1, x_2, \dots, x_n\}$ og fordelingsfunktion P . Så defineres *variansen* for X som

$$\begin{aligned} \text{Var}(X) &= \sigma^2 = E[(X - \mu)^2] \\ &= (x_1 - \mu)^2 \cdot P(X = x_1) + \cdots + (x_n - \mu)^2 \cdot P(X = x_n) \\ &= \sum_{i=1}^n (x_i - \mu)^2 \cdot P(X = x_i). \end{aligned}$$

Tallet $\sigma = \sqrt{\text{Var}(X)}$ kaldes for spredningen af X .

Eksempel 1.3. Lad os fortsætte eksemplet med terningkastet. Vi vil bestemme middelværdien $\mathbb{E}(X)$ for den stokastiske variabel X , der beskriver et terningkast. Dette bestemmes ifølge definitionen som

$$\mathbb{E}(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{21}{6} = 3.5,$$

hvilket er tæt på det vi eksperimentelt havde bestemt gennemsnittet til at være. Variansen er så

$$\begin{aligned}\text{Var}(X) &= (1 - 3.5)^2 \cdot \frac{1}{6} + (2 - 3.5)^2 \cdot \frac{1}{6} + (3 - 3.5)^2 \cdot \frac{1}{6} \\ &\quad + (4 - 3.5)^2 \cdot \frac{1}{6} + (5 - 3.5)^2 \cdot \frac{1}{6} + (6 - 3.5)^2 \cdot \frac{1}{6} \approx 2.92\end{aligned}$$

Spredningen er derfor

$$\sigma \approx \sqrt{2.92} \approx 1.71.$$

Det er svært at give en god fortolkning af spredningen ud over at det er den gennemsnitlige afstand et slag vil have fra middelværdien.

Middelværdi og spredning af binomialfordeling

Vi husker på, at vi skriver, at en binomialfordelt stokastisk variabel X med sandsynlighedsparameter p og antalsparameter n som

$$X \sim B(n, p).$$

Sætning 2.1. *Lad X være en binomialfordelt stokastisk variabel - mere præcist $X \sim \text{Bin}(n, p)$. Så gælder der, at middelværdien μ og spredningen σ er givet ved*

$$\mu = n \cdot p, \quad \text{og} \quad \sigma = \sqrt{n \cdot p \cdot (1 - p)}.$$

Der gælder altså, at $\mathbb{E}[X] = n \cdot p$ og $\text{Var}[X] = n \cdot p \cdot (1 - p)$.

Vi vil ikke bevise denne sætning, men den kan bevises ved at udnytte, at fordelingsfunktionen for X er givet som

$$P(X = k) = \binom{n}{k} (p)^k (1 - p)^{n-k}.$$

Eksempel 2.2. Vi lader X beskrive den binomialfordelte stokastiske variabel der tilsvarende antallet af slåede sekser efter ti slag med en terning. I dette tilfælde har vi antalsparameter $n = 10$ og sandsynlighedsparameter $p = \frac{1}{6}$, og

$$X \sim \text{Bin}(10, \frac{1}{6}).$$

Middelværdien af X er altså

$$\mathbb{E}[X] = 10 \cdot \frac{1}{6} \approx 1.66.$$

Vi forventer derfor at få omtrent 1,6 seksere på ti slag. Variansen af X er tilsvarende

$$\text{Var}[X] = 10 \cdot \frac{1}{6} \cdot \frac{5}{6} = \frac{50}{36} \approx 1.3$$

Betydning af spredning

Hvis vi modtager et datasæt, og vi forventer, at dette data er skabt af en binomialfordeling (eller en hvilken som helst anden fordeling), så vil det være oplagt at bruge spredningen og middelværdien til at bestemme, om det er sandsynligt, at dette data følger en binomialfordeling. Vi vil derfor gerne bestemme sandsynligheden for at ligge inden for den gennemsnitlige afvigelse for en binomialfordeling. Vi betragter igen en binomialfordelt stokastisk variabel

$$X \sim B(n, p).$$

Sandsynligheden for at X ligger inden for den gennemsnitlige afvigelse σ fra middelværdien μ må være

$$P(\mu - \sigma \leq X \leq \mu + \sigma).$$

Tilsvarende kan vi bestemme sandsynligheden for at ligge inden for to eller tre gennemsnitlige afvigelser som

$$P(\mu - 2 \cdot \sigma \leq X \leq \mu + 2 \cdot \sigma), \text{ og } P(\mu - 3 \cdot \sigma \leq X \leq \mu + 3 \cdot \sigma).$$

Disse sandsynligheder afhænger selvfølgelig af, hvad fordelingen af X er.

Eksempel 3.1. Lad X være en binomialfordelt stokastisk variabel

$$X \sim \text{Bin}(1000, 0.1),$$

der beskriver behandlingen af 1000 patienter med et lægemiddel, der har en succesrate på 10%. Middelværdien for X vil så være

$$\mathbb{E}[X] = \mu = 0,1 \cdot 1000 = 100,$$

og spredningen vil være

$$\sqrt{\text{Var}[X]} = \sigma = \sqrt{1000 \cdot 0.1 \cdot 0.9} = \sqrt{90} \approx 9.5.$$

Sandsynligheden for at ligge inden for en gennemsnitlig afvigelse vil så være

$$P(100 - 9,5 \leq X \leq 100 + 9,5) = 0.684.$$

Sandsynligheden for at ligge inden for en gennemsnitlig afvigelse fra middelværdien er derfor 68% (dette tal er bestemt med en computer). Tilsvarende kan vi finde sandsynligheden for at ligge inden for to eller tre gennemsnitlige afvigelser som

$$P(100 - 2 \cdot 9,5 \leq X \leq 100 + 2 \cdot 9,5) = 0.950,$$

$$P(100 - 3 \cdot 9,5 \leq X \leq 100 + 3 \cdot 9,5) = 0.997.$$

Derfor er sandsynligheden for at ligge inden for to spredninger fra middelværdien 95% og inden for tre spredninger fra middelværdien 99.7%. Sandsynligheden for at mindre end 70 personer helbredes er altså forsvindende lille, hvis medicinalfirmaet taler sandt. Tilsvarende vil sandsynligheden for at mere end 130 personer helbredes også være forsvindende lille.

Opgave 1

- i) Slå plat og krone, og lad $X = 1$, hvis udfaldet er plat og $X = 0$ ellers. Hvad er sredningen af X ?
- ii) Slå plat og krone to gange og lad X være antallet af plat. Hvad er spredningen af X ?
- iii) Kast en 12-sidet terning og lad X være antallet af øjne. Hvad er spredningen af X ?

Opgave 2

Fordelingen for en stokastisk variabel X kan beskrives ved følgende tabel.

x	-5	2	3	4	7	11
$P(X = x)$	0.1	0.2	0.1	0.4	0.05	0.15

- i) Bestem middelværdien for X .
- ii) Bestem spredningen for X .

Opgave 3

- i) Lad X være en Bernoulli-fordelt stokastisk variabel med sandsynlighedsparameter p . Hvad er middelværdien og variansen af X ?
- ii) Forklar med ord, hvorfor det giver god mening, at middelværdien for en binomialfordelt stokastisk variabel er $n \cdot p$.

Opgave 4

- i) Et lægemiddel helbreder 15% af behandlede personer. Lad X beskrive antallet af helbredte personer, når vi prøver at helbrede 10000 personer. Hvad er middelværdien og variansen af X ?
- ii) Vi slår fem gange med en terning, og lader X beskrive antallet af gange, vi slår mere end 4. Hvad er middelværdien og variansen for X ?

Opgave 5

- i) Lad $X \sim \text{Bin}(100, 1/3)$. Hvad er spredningen for X ? Hvad er sandsynligheden for at ligge inden for en spredning fra middelværdien? Hvad med to spredninger?

- ii) Hvad er sandsynligheden for, at få mindre end 20 ettere, hvis vi slår med en terning 100 gange?

Opgave 6

En producent af en bestemt type bildæk lover, at kun 0,01% af deres dæk er defekte fra produktionen. En importør af disse dæk synes, at lidt for mange af dækkene er defekte. Han har importeret 1.200 dæk, og 5 af disse var defekte. Hvad er sandsynligheden for at mere end 4 dæk er defekte? Har importøren grund i hans mistanke.