Betydning af spredning

Middelværdi og varians af binomialfordeling

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

$$X \sim \text{Bin}(n, p).$$

Sætning 1.1. *Lad X være en binomialfordelt stokastisk variabel - mere præcist $X \sim \text{Bin}(n, p)$. Så gælder der, at*

$$\mathbb{E}(X) = n \cdot p, \quad \text{og} \qquad \text{Var}(X) = n \cdot p \cdot (1 - p).$$

Der gælder altså, at $\mu = n \cdot p$ og $\sigma^2 = n \cdot p \cdot (1-p)$. Der gælder derfor, at spredningen σ for X er givet som

$$\sigma = \sqrt{n \cdot p \cdot (1 - p)}.$$

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 1.2. Vi lader X beskrive den binomialfordelte stokastiske variabel der tilsvarer 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

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

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

$$X \sim \text{Bin}(n, p).$$

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

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

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

Eksempel 2.1. 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.

Parameterestimation

Vi vil introducere parameterestimation med et eksempel.

Eksempel 3.1. Jeg stiller en opgave bestående af 10 delopgaver, og jeg stiller opgaverne, så det er tilfældigt om en opgave bliver besvaret rigtigt eller forkert. Et rigtigt svar er lige sandsynligt for alle opgaverne, og så tæller jeg, hvor mange opgaver, I har fået rigtige hver i sær. Jeg vil gerne vide, hvad sandsynligheden for et rigtigt svar er. Følgende frekvens af rigtige svar er bestemt:

Antal rigtige opgaver	0	1	2	3	4	5	6	7	8	9	10
Frekvens	0	0	0	0	2	1	3	5	5	2	2

Vi husker på, at middelværdien for binomialfordelingen var $n \cdot p$. Vi kender n, den er 10, og vi ved, hvordan vi bestemmer et estimat for middelværdien:

$$\hat{\mu} = 0 \cdot 0 + 1 \cdot 0 + 2 \cdot 0 + 3 \cdot 0 + 4 \cdot \frac{2}{20} + 5 \cdot \frac{1}{20} + 6 \cdot \frac{3}{20} + 7 \cdot \frac{5}{20} + 8 \cdot \frac{5}{20} + 9 \cdot \frac{2}{20} + 10 \cdot \frac{2}{2} = \frac{36}{5} = n \cdot \hat{p},$$

hvor $\hat{p}$ betegner vores estimat for sandsynligheden for succes. Vi kan så bestemme dette estimat som

$$\frac{\frac{36}{5}}{10} \approx 0.72 = \hat{p}$$

Derfor er det bedste bud på sandsynligheden for et korrekt svar givet ved 0.72.

Opgave 1

i) Et medicinsk præparat 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 2

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 3

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.