Konfidensintervaller

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

$$X \sim \mathrm{B}(n, p).$$

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

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

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

Eksempel 1.1. Lad X være en binomialfordelt stokastisk variabel

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

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

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

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.

Konfidensintervaller

Sætning 2.1. *Har vi en binomialfordelt stikprøve på n elementer med sandsynlighedsparameter p, så er et 95% konfidensinterval givet ved*

$$\left[\hat{p} - 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{p} + 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right],$$

hvor $\hat{p}$ er stikprøveestimatet for p.

Dette skal forstås på følgende måde. Laver vi en stikprøve på n elementer 100 gange, så vil vi få 100 forskellige konfidensintervaller. Så vil vi forvente at den rigtige p vil ligge inde i 95 af disse 100 konfidensintervaller.

Eksempel 2.2. I en meningsmåling fra Voxmeter er 1000 personer blevet spurgt, hvad de vil stemme til det danske folketingsvalg, og vi antager, at personerne er en repræsentativ stikprøve for den danske befolkning. 300 personer har svaret, at de vil stemme på Socialdemokratiet. Vores stokastiske eksperiment har bestået i at spørge 1000 personer, og en succes har været svaret: "Socialdemokratiet". Vi bestemmer nu et 95% konfidensinterval for denne stikprøve. Vores estimat på $\hat{p}$ må være

$$\hat{p} = \frac{300}{1000} = 0.3.$$

Usikkerheden eller fejlmarginen er så

$$2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 2\sqrt{\frac{0.3 \cdot 0.7}{1000}} = 0.0289 = 2.89\%.$$

Konfidensintervallet for stikprøven bliver da

$$\left[\hat{p} - 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] = [0.3 - 0.0289, 0.3 + 0.0289]$$
$$= [0.2710, 0.3289].$$

Den rigtige sandsynlighedsparameter p må være andelen af personer, der stemmer på Socialdemokratiet til folketingsvalget (Alt efter hvad vi mener, vi måler på). Derfor vil socialdemokratiet hvis befolkningen ikke skifter mening mellem meningsmålingen og folketingsvalget med 95% sikkerhed have en vælgertilslutning på mellem 27.10% og 32.89%. Socialdemokraterne fik som bekendt 27.5% af stemmerne, hvilket er inden for den statistiske usikkerhed.

Eksempel 2.3. Vi bestemmer et 95% konfidensinterval for Dansk Folkeparti. Deres tilslutning i meningsmålingen var 25 personer. Derfor er stikprøveandelen

$$\hat{p} = \frac{25}{1000} = 0.025 = 2.5\%.$$

Den statistiske usikkerhed er derfor

$$2\sqrt{\frac{\hat{p}(1-\hat{p})}{1000}} = 0.0099 = 0.99\%.$$

Konfidensintervallet lyder så

$$[0.025 - 0.0099, 0.025 + 0.0099] = [0.0151, 0.0349].$$

Tilslutningen til DF vil altså med 95% sikkerhed ligge mellem 1.51% og 3.49%. De er altså ikke sikre på at ligge over spærregrænsen på 2%. DF fik til valget 2.6% af stemmerne, hvilket er inden for den statistiske usikkerhed.

Opgave 1

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 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.

Opgave 4

Et lægemiddel helbreder 5 ud af 100 personer.

i) Bestem et 95%-konfidensinterval for andelen af antallet af helbredte.

ii) Producenten af lægemiddelet påstår, at 12% af de behandlede med lægemiddelet bliver helbredt. Ligger dette udfald inden for den statistiske usikkerhed?

Opgave 5

Til det seneste valg spurgte vi 300 personer om, hvad de stemte. 30 personer stemte på liste Q.

i) Bestem et 95%-konfidensinterval for andelen af stemmer til liste Q.

ii) Til næste valg får partiet 8% af stemmerne. Ligger dette tal inden for den statistiske usikkerhed? Kan vi konkludere, at tilslutningen til partiet er gået tilbage?

Opgave 6

Vi kaster med en mønt 150 gange og får plat 70 gange.

i) Bestem et 95%-konfidensinterval for andelen af plat ved kast med mønt

ii) Kan vi sige noget om, hvorvidt denne mønt er fair?

Opgave 7

En mand finder 3 sten i sit glas med 70 oliven og synes, at det er lidt for mange. Producenten har jo lovet, at glasset er 99% stenfri.

i) Bestem et 95%-konfidensinterval for andelen af sten i glasset.

ii) Har manden grund i sin mistanke om, at der er for mange sten i glasset?