Middelværdi og spredning fra stikprøve

Estimation af parametre

I de eksempler, vi har arbejdet med, har vi antaget, at middelværdien og spredningen er kendte værdier, men dette vil kun meget sjældent være tilfældet i normalfordelte stokastiske variable, der beskriver virkelige fænomener. Vi skal derfor have en måde at estimere middelværdien μ og spredningen σ for den underliggende normalfordelte stokastiske variabel X, vi forventer kan beskrive vores stikprøve. Middelværdien for en stokastisk variabel er et slags ikke-realiseret gennemsnit, og derfor estimerer vi den ved at bestemme gennemsnittet af datasættet.

Definition 1.1 (Gennemsnit og spredning)**.** Lad $x_1, x_2, \ldots, x_n$ være en stikprøve. Så defineres middelværdiestimatet $\hat{\mu}$ som gennemsnittet af stikprøven

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

Spredningsestimatet $\hat{\sigma}$ defineres som kvadratroden af den gennemsnitlige kvadratiske variation fra middelværdien

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

Til tider bruges også spredningsestimatet s givet ved

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \hat{\mu})^2}{n - 1}}.$$

Dette er i princippet et bedre bud på den *rigtige* spredning, men vi vil ikke komme præcist ind på hvorfor. Intuitionen er dog følgende; stikprøveværdierne x_i ligger generelt tættere på stikprøvemiddelværdien $\hat{\mu}$ end på den rigtige middelværdi μ, og derfor vil spredningen blive lidt for lille. Ved at betragte den gennemsnitlige afstand mellem $\hat{\mu}$ og μ kan man så komme frem til estimatet s.

Eksempel 1.2. Vi har følgende data, og vi ønsker at bestemme middelværdien og spredningen for datasættet. Dette gøres i Maple ved at skrive

```
with(Gym):
middel(Data)
spredning(Data)
```

Vi gør dette i Maple og får, at middelværdien er

$$\hat{\mu} = 198.97$$

og spredningen er

$$\hat{\sigma} = 15.19.$$

Er data normalfordelt?

Vi vil gerne kunne afgøre, om et datasæt er normalfordelt. Til dette laver vi det, vi kalder et *fraktilplot*. Først skal vi dog bruge en sætning, der siger noget om, hvordan vi transformerer en normalfordelt stokastisk variabel til en standardnormalfordelt stokastisk variabel.

Sætning 2.1. *Lad $X \sim N(\mu, \sigma)$ være en normalfordelt stokastisk variabel og lad*

$$Z = \frac{X - \mu}{\sigma}.$$

Så gælder der, at

$$Z \sim N(0, 1).$$

Bevis. Vi betragter fordelingsfunktionen for Z:

$$\begin{aligned} P(Z < a) &= P(\frac{X-\mu}{\sigma} \leq a) \\ &= P(X - \mu < a\sigma) \\ &= P(X < a\sigma + \mu) \\ &= \int_{-\infty}^{a\sigma+\mu} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx. \end{aligned}$$

Vi laver nu substitutionen

$$z = \frac{x - \mu}{\sigma},$$

og får, at

$$\frac{dz}{dx} = \frac{x}{\sigma}$$

og dermed, at

$$dz\sigma = dx.$$

For at lave substitutionen skal vi tilsvarende ændre grænserne. Den nedre grænse giver

$$\frac{-\infty - \mu}{\sigma} = -\infty,$$

og den øvre grænse giver

$$\frac{a\sigma + \mu - \mu}{\sigma} = a.$$

Vi substituerer

$$\begin{aligned}
\int_{-\infty}^{a\sigma+\mu} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx &= \int_{-\infty}^{a} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz\sigma \\
&= \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz \\
&= \int_{-\infty}^{a} \varphi(z) dz \\
&= \Phi(a).
\end{aligned}$$

Da Z har Φ som fordelingsfunktion, så må Z være standardnormalfordelt. ■

Vi ved, at Φ er en voksende funktion. Den har derfor en invers funktion Φ^{-1}. Da $X \sim N(\mu, \sigma)$ har fordelingsfunktion

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

så må der gælde, at

$$z = \Phi^{-1}(F(x)) = \Phi^{-1}\left(\Phi\left(\frac{x-\mu}{\sigma}\right)\right) = \frac{x-\mu}{\sigma}.$$

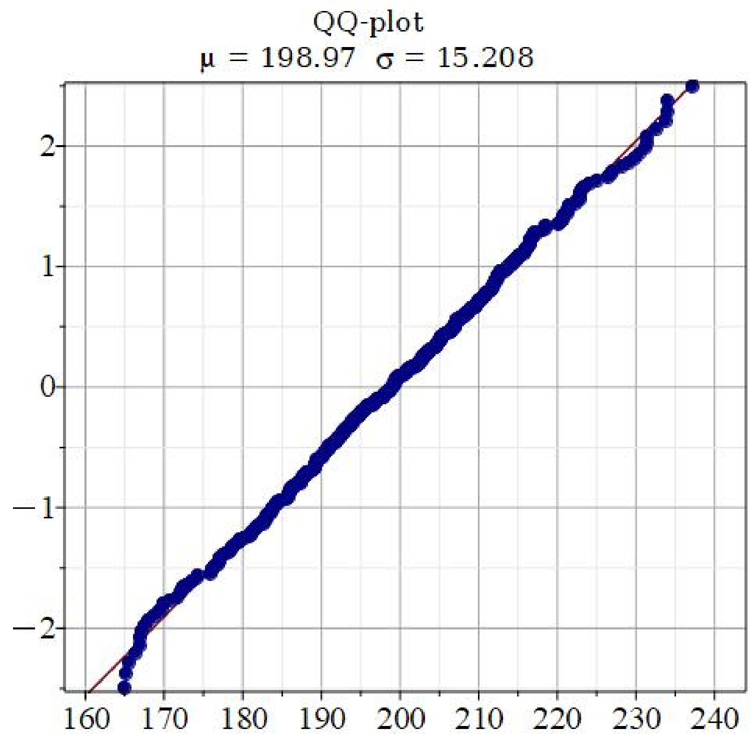
Vi kan altså teste om et datasæt er normalfordelt ved at plotte dataet op mod de teoretiske z-værdier. Følger disse den rette linje

$$z = \frac{x-\mu}{\sigma},$$

så vil datasættet forventes at være normalfordelt. Dette er intuitionen bag fraktilplottet. Det laves i Maple ved at skrive

```
with(Gym):
QQplot(Data)
```

Eksempel 2.2. Vi har lavet et fraktilplot på datasættet fra før i Maple. Dette kan ses af Fig. 1.



Figur 1: Fraktilplot for datasæt

Dataet ligger pænt langs med den rette linje $z = (x - 198.97)/15.208$, og vi antager derfor, at det er normalfordelt.

Opgave 1

Dette data beskriver IQ for 213 værnepligtige.

i) Bestem middelværdi og spredning for IQ af de værnepligtige.

ii) Afgør, om dataen kan antages at være normalfordelt.

iii) Bestem sandsynligheden for at have en IQ på under 130.

iv) Bestem det tal, så 99% af befolkningen har en IQ på mindre end x.

Opgave 2

173 raske personer har fået målt deres temperatur. Resultatet er i dette datasæt.

i) Bestem middelværdi og spredning for temperaturen.

ii) Afgør, om dataen kan antages at være normalfordelt.

iii) Bestem sandsynligheden for at have en temperatur på under 40 grader.

iv) Bestem sandsynligheden for at have en temperatur mellem 35 og 38 grader.

Opgave 3

501 kvinder har målt deres højde. Resultatet kan findes her.

1. Bestem middelværdi og spredning for højden.
2. Afgør, om dataen kan antages at være normalfordelt.
3. Bestem sandsynligheden for, at en kvinde er mindre end 150 cm
4. Bestem et 95%-konfidensinterval for den rigtige middelværdi μ.