Hvis du fisker torsk på Vestfjorden ligger det en en stor sannsynlighetsmodell i bånn, ofte normalfordelingen, som forteller hva slags torsk du kan få. Normalfordelingen er bestemt av μ og σ. Disse kjenner vi gjerne ikke, men ved å veie noen lofottorsk, kan vi gjennom statistiske metoder estimere dem.

Utvalg

Anta at en lofottorskstamme er normalfordelt med $\mu = 7$ kg og $\sigma = 1.5$ kg. Vi plukker ut ni tilfeldige lofottorsk. De veier (i kg)

8.4 5.7 7.0 6.4 9.3 8.1 8.6 9.1 8.0

Hva sier disse tallene oss om torskestammen? Gjennomsnittsvekten er 7.06 kg. Dersom torskestammen er normalfordelt med $\mu = 7.0$ kg, betyr det at hvis vi veier alle torskene i stammen og deler på antall veide torsk, skal vi få *akkurat* 7 kg. Dette settet med ni torsk veide ikke nøyaktig 7 kg i gjennomsnitt, og det vil vi heller ikke få med mindre vi veier alle torsk i stammen.
Disse ni torskene kalles et *utvalg*. Det er som sagt naturlig at torskene i et utvalg ikke har gjennomsnittsvekt på nøyaktig 7 kg. Men om et tilfeldig utvalg har betydelig høyere eller lavere gjennomsnitt enn 7 kg, kan det være grunn til å tvile på om normalfordelingen til torskevekten virkelig virkelig er sentrert rundt $\mu = 7$ kg. Sannsynlighetsmodellen for hvor mye gjennomsnittsvekten på utvalget avviker fra μ, kan nemlig bestemmes nøyaktig, og dette benytter vi oss av i statistikk.

Estimator for μ

Som regel estimerer vi μ, og estimatoren kalles som regel $\hat{\mu}$. Det kan gjøres på flere måter. En klassisk estimator er

$$\hat{\mu} = \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n},$$

gjennomsnittet av n målinger trukket fra den samme fordelingen.

Eksempel. Gjennomsnittsvekten 7.06 kg på utvalget er et estimat for forventningsverdien μ. △

Nå er også $\overline{X}$ en stokastisk variabel, siden den er en sum av stokastiske variable. Hvis vi regner ut forventningen til $\overline{X}$, får vi

$$E(\overline{X}) = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{\sum_{i=1}^{n} E(X_i)}{n} = \frac{n\mu}{n}.$$

Likningen

$$E(\overline{X}) = \mu$$

forteller oss at dersom vi veier mange lofottorsk, vil forventet gjennomsnittsvekt i utvalget være μ, og vi sier at estimatoren $\hat{\mu}$ er *forventningsrett*.

Estimator for μ

Hvis vi skal estimere σ må vi bruke formelen

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

Det går an å vise at denne formelen er forventningsrett, men det skal ikke vi gjøre. Dersom man deler på n istedet for $n-1$, får man *ikke* en forventningsrett estimator.

Konfidensintervall - kjent σ

La oss tenke at du måler en parameter μ med et måleapparat som er grundig testet i laboratorium, slik at avviket σ er kjent - dersom du gjør en måling $\hat{\mu}$ med apparatet, kan fabrikanten fortelle deg nøyaktig hvor presis denne målingen er. Som oftest er målingen $\hat{\mu}$ normalfordelt rundt den faktiske verdien μ.

Eksempel. Promillen til en person med promille på $\mu = 0.8$ måles med et normalfordelt måleinstrument der $\sigma = 0.1$. Sannsynligheten for at måleapparatet estimerer en promille på mellom 0.7 og 0.9 er

$$P(0.7 < \hat{\mu} < 0.9) = P(\mu - \sigma < \hat{\mu} < \mu + \sigma) = 0.6827.$$

△

I situasjoner der noe estimeres, skal man tenke at estimatet gir et intervall der det er en viss sannsynlighet for at den faktiske parameteren befinner seg - *et konfidensintervall.*

Eksempel. Vi måler promillen til samme person som i sted, og får $\hat{\mu} = 0.87$. Det kan vises at

$$P(0.77 < \mu < 0.97) = P(\hat{\mu} - \sigma < \mu < \hat{\mu} + \sigma) = 0.6827,$$

men det skal ikke vi gjøre. Det er altså 68% sannsynlighet for at mannens promille μ ligger i intervallet $(0.77, 0.97)$, og vi sier at dette er et 68% konfidensintervall for μ. △

Dersom vi gjør flere målinger, blir konfidensintervallet mer presist.

Eksempel. Vi måler promillen til samme person som i sted fem ganger, og tar gjennomsnittet av målingene. Hvis X_i er en stokastisk variabel som beskriver den i-te målingen, vil $\hat{\mu} = \overline{X}$. Regnereglene for varians gir at standardavviket til $\overline{X}$ blir $\sigma/\sqrt{5} = 0.045$. Vi får

$$P(0.755 < \hat{\mu} < 0.845) = 0.6827.$$

Intervallet $\hat{\mu}$ havner i med 68% sannsynlighet er kortere når vi tar gjennomsnittet av flere målinger. △

Konfidensintervall - ukjent σ