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

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

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*.

Eksempel. En jeg kjenner, er overtroisk. Han legger ekstra vekt på den første målingen, og derfor beregner han gjennomsnitt med formelen

$$\hat{\mu} = \frac{1}{n}\left(2X_1 + X_2 + ... + X_n\right).$$

Dette er ikke en forventningsrett estimator, for

$$E(\hat{\mu}) = \frac{n+1}{n}\mu.$$

Den skyter konsekvent litt over. △

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

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

△

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 målingene mer presise.

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. △

Dersom vi gjør flere målinger, blir også konfidensintervallet trangere.

Eksempel. Vi målte promillen fem ganger og fikk 0.82, 0,79, 0.83, 0.83 og 0.80. Da blir

$$\hat{\mu} = \overline{X} = \frac{0.82 + 0.79 + 0.83 + 0.83 + 0.80}{5} = 0.814.$$

Et 95% konfidensintervall for den faktiske promillen μ blir

$$(\hat{\mu} - 1.96\sigma/\sqrt{5} < \mu < \hat{\mu} + 1.96\sigma/\sqrt{5}) = (0.73, 0.9).$$

△

Vi setter opp denne prosedyren som et teorem.

Teorem 6.1. *Dersom μ estimeres med gjennomsnittet $\hat{\mu} = \overline{X}$ av n målinger, og σ er kjent for hver måling, er det 95% sannsynlighet for at*

$$\hat{\mu} - 1.96\sigma/\sqrt{n} < \mu < \hat{\mu} + 1.96\sigma/\sqrt{n}.$$

Konfidensintervall - ukjent σ

Dersom μ skal estimeres uten at σ er kjent, må σ estimeres. Når man har estimert σ, bruker man t-fordelingen istedet for normalfordelingen. Denne fordelingen ligner på normalfordelingen, og oppfører seg likt, men er litt flatere. I tabellens venstre kolonne angis hvor mange målinger som er gjort (m), og i raden på toppen er forskjellige sannsynligheter på formen

$$P(T > a\hat{\sigma}/\sqrt{m})$$

og så gir tabellen antall estimerte standardavvik a.

Eksempel. Vi kaster en terning ti ganger, og får

x_1	3
x_2	2
x_3	5
x_4	6
x_5	1
x_6	4
x_7	4
x_8	3
x_9	5
x_{10}	4

Gjennomsnittlig antall øyne på disse ti kastene er $\hat{\mu} = \overline{X} = 3.7$, og estimert standardavvik er $\hat{\sigma} = 1.42$. Siden det er ti kast, må vi slå opp på $m = 10$ i t-fordelingen, og der står det for eksempel at

$$P(T > 1.372 \cdot 1.42/\sqrt{10}) = P(T > 0.62) = 0.10.$$

Dette betyr at det er 10% △

m \ α	0.20	0.10	0.05	0.025	0.02	0.01	0.005
1	1.376	3.078	6.314	12.706	15.894	31.821	63.656
2	1.061	1.886	2.920	4.303	4.849	6.965	9.925
3	0.978	1.638	2.353	3.182	3.482	4.541	5.841
4	0.941	1.533	2.132	2.776	2.999	3.747	4.604
5	0.920	1.476	2.015	2.571	2.757	3.365	4.032
6	0.906	1.440	1.943	2.447	2.612	3.143	3.707
7	0.896	1.415	1.895	2.365	2.517	2.998	3.499
8	0.889	1.397	1.860	2.306	2.449	2.896	3.355
9	0.883	1.383	1.833	2.262	2.398	2.821	3.250
10	0.879	1.372	1.812	2.228	2.359	2.764	3.169
11	0.876	1.363	1.796	2.201	2.328	2.718	3.106
12	0.873	1.356	1.782	2.179	2.303	2.681	3.055
13	0.870	1.350	1.771	2.160	2.282	2.650	3.012
14	0.868	1.345	1.761	2.145	2.264	2.624	2.977
15	0.866	1.341	1.753	2.131	2.249	2.602	2.947
16	0.865	1.337	1.746	2.120	2.235	2.583	2.921
17	0.863	1.333	1.740	2.110	2.224	2.567	2.898
18	0.862	1.330	1.734	2.101	2.214	2.552	2.878
19	0.861	1.328	1.729	2.093	2.205	2.539	2.861
20	0.860	1.325	1.725	2.086	2.197	2.528	2.845
21	0.859	1.323	1.721	2.080	2.189	2.518	2.831
22	0.858	1.321	1.717	2.074	2.183	2.508	2.819
23	0.858	1.319	1.714	2.069	2.177	2.500	2.807
24	0.857	1.318	1.711	2.064	2.172	2.492	2.797
25	0.856	1.316	1.708	2.060	2.167	2.485	2.787
26	0.856	1.315	1.706	2.056	2.162	2.479	2.779
27	0.855	1.314	1.703	2.052	2.158	2.473	2.771
28	0.855	1.313	1.701	2.048	2.154	2.467	2.763
29	0.854	1.311	1.699	2.045	2.150	2.462	2.756
30	0.854	1.310	1.697	2.042	2.147	2.457	2.750
31	0.853	1.309	1.696	2.040	2.144	2.453	2.744
32	0.853	1.309	1.694	2.037	2.141	2.449	2.738
33	0.853	1.308	1.692	2.035	2.138	2.445	2.733
34	0.852	1.307	1.691	2.032	2.136	2.441	2.728
35	0.852	1.306	1.690	2.030	2.133	2.438	2.724
36	0.852	1.306	1.688	2.028	2.131	2.434	2.719
37	0.851	1.305	1.687	2.026	2.129	2.431	2.715
38	0.851	1.304	1.686	2.024	2.127	2.429	2.712
39	0.851	1.304	1.685	2.023	2.125	2.426	2.708
40	0.851	1.303	1.684	2.021	2.123	2.423	2.704
50	0.849	1.299	1.676	2.009	2.109	2.403	2.678
60	0.848	1.296	1.671	2.000	2.099	2.390	2.660
70	0.847	1.294	1.667	1.994	2.093	2.381	2.648
80	0.846	1.292	1.664	1.990	2.088	2.374	2.639
∞	0.842	1.282	1.645	1.960	2.054	2.326	2.576

Sannsynlighetstettheten til t-fordelingen er

$$f(t) = \frac{1}{K}\left(1 + \frac{t^2}{n-1}\right)^{-n/2},$$

der

$$K = \int_{-\infty}^{\infty}\left(1 + \frac{t^2}{n-1}\right)^{-n/2} dx.$$

Denne skal vi ikke bruke til noe, men det er jo artig å vite. Her er et plot.