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 statistikken.

Estimatorer

Gjennomsnittsvekten 7.06 kg er altså et estimat for forventningsverdien $\mu = 7$ kg. Vi estimerer alltid forventningen ved å ta gjennomsnittet av målingene. De n målingene kaller vi x_i, og skriver

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Vi bruker 'hatt' for indikere estimering , og $\hat{\mu}$ kalles en **estimator** for μ. Merk at hver enkelt x_i er verdien til en stokastisk variabel, siden den enkelte torskens vekt er normalfordelt. Men da er også $\hat{\mu}$ en vanlig stokastisk variabel, siden den er en sum av stokastiske variable. Hvis vi regner ut forventningen til $\hat{\mu}$, får vi

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

Likningen

$$E(\hat{\mu}) = \mu$$

forteller oss at dersom vi veier mange lofottorsk, vil gjennomsnittet krype mot μ, og vi sier at estimatoren $\hat{\mu}$ er **forventningsrett**.
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.

Hypotesetesting

Hypotesetest er en viktig statistisk metode for å analysere store sannsynlighetsmodeller. Her er oppskriften.

1. Sett opp en hypotese.
2. Kjør forsøk.
3. Analyser resultatene fra forsøket, og vurder om hypotesen din er forenlig med resultatene du fikk.

Nå skal vi opparbeide et rammeverk for statistisk analyse av normalfordelte størrelser. Det er hensiktsmessig å dele inn i to tilfeller: kjent σ, og ukjent σ.

Kjent σ

Dersom man måler promille hos en dame som kjører vinglete, er det med et måleinstrument som vil gi litt variasjon i promille fra måling til måling. Siden dette måleinstrumentet skal brukes til å dømme eller frikjenne folk etter straffeloven, er det grundig testet i lab før det blir brukt, og dermed kan vi anta at $\sigma = 0.07$ er kjent; det står i brukermanualen. Vi kjører nå en hypotesetest med en enkelt måling av promillegrensen. Vi følger oppskriften over.

1. Promillegrensen i Norge er 0.20, så vi setter opp som hypotese at damen har promille på 0.20. I så fall vil en enkelt måling være normalfordelt med $\mu = 0.20$ og $\sigma = 0.07$.
2. Så måler politiet promillen. De fikk at den ble 0.29.
3. Sannsynligheten for at måleinstrumentet viser 0.29 eller mer *dersom damens promille er 0.20* kan regnes ut til (*siden måleresultatet er normalfordelt*)

   $$P(x > 0.29) = P(x > \mu + 0.09) = P(x > \mu + 1.29\sigma) =$$

   $$1 - P(x < \mu + 1.29\sigma) = 1 - 0.9015 = 0.0985.$$

   Dette betyr at dersom damen har 0.2 i promille, er det hele 9.85% sannsynlighet for at apparatet viser mer enn 0.29 pga målefeil.

Tenk litt over dette eksemplet. Dersom vi skal si at verdien som avleses på måleinstrumentet er hennes faktiske promille, er det jo hele 50% sannsynlighet for at damen blir dømt ($P(x > \mu) = 0.50$), selv om hun skulle ha lovlig promille på 0.20. Dersom vi krever at apparatet skal vise 2.9 eller høyere for at hun skal dømmes, er det 9.85% sannsynlighet for at hun blir uskyldig dømt. Dersom vi krever at det kun skal være 5% sannsynlighet for justismord, må vi kreve at

$$P(x > \mu + a\sigma) = 0.05.$$

Vi ser i tabellen at dersom $a = 1.65$, har vi at

$$P(x < \mu + 1.65\sigma) = 0.95,$$

slik at

$$P(x > \mu + 1.65\sigma) = 0.05,$$

Dette betyr at vi må kreve at apparatet viser

$$\mu + 1.65\sigma = 0.2 + 1.65 \cdot 0.07 = 0.3155$$

før det er kun 5% prosent sannsynlighet for justismord. Dersom vi krever at det skal være kun 5% sannsynlighet for justismord, sier vi at hypotesetesten er gjennomføres med 5% **signifikans**.

Men 3.155 er ganske høyt over 2.0. Det kan vi gjøre noe med ved å ta flere målinger. La oss ta fem målinger istedet, og så bruke gjennomsnittet av dem som estimator for damens promille μ. Merk at når vi brukte en måling, var den målingen en estimator for μ. Men nå skal vi altså måle fem ganger. Den nye variabelen blir

$$\hat{\mu} = \frac{\sum_{i=1}^{5} x_i}{5},$$

der x_i er måling nr. i. Måling nr. i er en normalfordelt variabel med μ (hypotesens antagelse) og σ (kjent fra lab). Da vil $\hat{\mu}$ være en normalfordelt variabel med forventning μ (fordi den er forventningsrett), og vi kan, vha. regneregler for standardavvik (kap. 1.7), regne ut

$$\sigma(\hat{\mu}) = \frac{\sigma}{\sqrt{n}} = \frac{0.07}{\sqrt{5}} = 0.031$$

Vi ser altså at effekten av å måle flere ganger er å minke effektivt standardavvik. Hvis vi bruker det nye standardavviket, og krever at

$$P(\hat{\mu} > \mu + a\sigma(\hat{\mu})) = 0.05$$

gir jo normalfordelingstabellen fremdeles at dette er tilfellet dersom $a = 1.65$, men nå er

$$\mu + a\sigma(\hat{\mu}) = 0.2 + 1.65 \cdot 0.031 = 0.25.$$

Altså: dersom *gjennomsnittet* av de 5 målingene er over 0.25 er det nå kun 5% sannsynlighet for justismord dersom hun dømmes på grunnlag av målingene.

Ukjent σ

Anta at en fyr som er veldig stolt av å være fra Lofoten påstår lofottorskstammen er normalfordelt med $\mu = 10$ kg, men at det ikke finnes noen statistikkeksperimenter til nå som underbygger dette. Da må vi prøve å finne det ut, siden han som sa det var en sykt irriterende fyr. Vi bruker tallene fra torskeeksemplet over. I dette tilfellet har vi ikke kjennskap til den ekte μ og σ. Derfor må vi ta til takke med tallene fra torskeeksemplet, og μ og σ må estimeres. Vi estimerer

$$\hat{\mu} = \frac{8.4 + 5.7 + 7.0 + 6.4 + 9.3 + 8.1 + 8.6 \ + 9.1 + 8.0}{9}$$
$$= 7.84 \text{ kg}.$$

og

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{9}(x_i - \hat{\mu})^2}{8}} = 1.22 \text{ kg}$$

Hypotesetesten gjennomføres nå som før, men siden vi estimerer σ, må vi bruke **t-fordelingen** istedet for normalfordelingen. **En normalfordelt størrelse med gitt μ og ukjent σ er t-fordelt med μ og $\hat{\sigma}$ som parametre, dersom σ estimeres med $\hat{\sigma}$.** Akkurat som med normalfordelingen, må sannsynligheter fra t-fordelingen finnes i en tabell. T-fordelingen ligner ganske mye på normalfordelingen; den bare er litt flatere. Også avhenger den av hvor mange målinger man har brukt for å estimere σ. Sannsynlighetene i t-fordelingen er gitt som følger: Ta antall observasjoner, og trekk fra en. I kolonnen til venstre, skal man finne tallet man fikk. Altså, dersom det er n observasjoner, skal man velge $n - 1$ i denne kolonnen. Tallene i den rekken man da har valgt, gir **hvor mange estimerte standardavvik $\hat{\sigma}$ over μ** man må være for at sannsynligheten

$$P(x > \mu + a\hat{\sigma})$$

skal være 0.1, 0.05, 0.025, 0.01, 0.005, 0. 001, eller 0.0005. T-fordelingen er symmetrisk, akkurat som normalfordelingen, slik at

$$P(x > \mu + a\hat{\sigma}) = P(x < \mu - a\hat{\sigma})$$

Nå fortsetter vi med torskeeksemplet.

prob	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
df							
1	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

1. Vi setter opp hypotesen at $\mu = 10$ kg, slik som lofotværingen påstår. Dette betyr at vi antar at torsken er normalfordelt med $\mu = 10$ kg.

2. Vi har funnet at $\hat{\mu} = 7.84$. Vi har estimert $\hat{\sigma} = 1.22$ med 9 observasjoner, og det betyr at vi skal bruke $n-1=8$ i tabellen. Der står det at med 9 observasjoner, har vi at
   - sannsynligheten er 0.1 for å havne mer enn 1.397 estimerte standardavvik nedenfor $\mu = 10$ kg
   - sannsynligheten er 0.05 for å havne mer enn 1.860 estimerte standardavvik nedenfor $\mu = 10$ kg
   - sannsynligheten er 0.025 for å havne mer enn 2.306 estimerte standardavvik nedenfor $\mu = 10$ kg
   - sannsynligheten er 0.01 for å havne mer enn 2.896 estimerte standardavvik nedenfor $\mu = 10$ kg
   - sannsynligheten er 0.005 for å havne mer enn 3.355 estimerte standardavvik nedenfor $\mu = 10$ kg
   - sannsynligheten er 0.001 for å havne mer enn 4.501 estimerte standardavvik nedenfor $\mu = 10$ kg
   - sannsynligheten er 0.0005 for å havne mer enn 5.041 estimerte standardavvik nedenfor $\mu = 10$ kg
3. Vi har jo at
$$\frac{\mu - \hat{\mu}}{\hat{\sigma}} = \frac{10 - 7.06}{1.48} = 2.41$$
så den observerte verdien er 2.41 estimerte standardavvik under antatt μ. Dette betyr at vi kan forkaste hypotesen $\mu = 10$ kg med 10% signifikans, med 5% signifikans, og med 2.5% signifikans, men ikke med 1% signifikans.

Sammenligning av to grupper

Nå skal vi sammenligne to normalfordelte variable x og y. Dette er nesten det samme som i sted; vi må bare jukke litt med μ og σ for at det skal gå i orden. Ideen er å studere differansen $x-y$, for å finne ut om x og y er forskjelige.

Vi tar et eksempel med to lofottorskstammer x og y. Vi antar at de er normalfordelte med *identiske* standardavvik σ. Ni og ti torsk av de to stammene ble veid, og resultatet ble

8.4 5.7 7.0 6.4 9.3 8.1 8.6 9.1 8.0

og

8.1 4.7 6.3 6.9 5.6 8.0 6.7 7.5 10.8 5.9

kg. Så er spørsmålet: Er disse tallene forenlige med at torskestammene har samme gjennomsnittsvekt? Som hypotese setter vi opp at $\mu_x - \mu_y = 0$, altså at torskestammene er identisk normalfordelte variable. Så beregner vi $\hat{\mu}_x$ og $\hat{\mu}_y$. De blir
$$\hat{\mu}_x = 7.84.$$
og
$$\hat{\mu}_y = 7.19.$$
Herfra må vi dele analysen opp i to biter, kjent og ukjent standardavvik.

Kjent σ

Hvis vi antar kjent standardavvik $\sigma = 1.5$ kg, kan vi bruke normalfordelingen. Dette er fordi at siden både x og y er normalfordelte variable, er også $\hat{\mu}_x$ og $\hat{\mu}_y$ normalfordelte variable med forventning μ_x og μ_y. Nå er vi interessert i variabelen
$$\hat{\mu}_x - \hat{\mu}_y,$$
og vi kan beregne
$$E(\hat{\mu}_x - \hat{\mu}_y) = E(\hat{\mu}_x) - E(\hat{\mu}_y) = \mu_x - \mu_y = 0.$$
Standardavvikene til variabelen $\hat{\mu}_x - \hat{\mu}_y$ kan beregnes ved formlene for standardavvik (kap. 1.7), og vi får
$$\begin{aligned}\sigma(\hat{\mu}_x - \hat{\mu}_y) &= \sqrt{\sigma^2(\hat{\mu}_x) + \sigma^2(\hat{\mu}_y)} \\ &= \sqrt{\frac{\sigma^2}{\sqrt{9}} + \frac{\sigma^2}{\sqrt{10}}} = \sigma\sqrt{\frac{1}{\sqrt{9}} + \frac{1}{\sqrt{10}}},\end{aligned}$$
der vi har brukt at det er samme kjente standardavvik σ på de to stammene. Nå er det bare å sette inn $\sigma = 1.5$, og da får vi
$$\sigma(\hat{\mu}_x - \hat{\mu}_y) = 1.5 \cdot \sqrt{\frac{1}{\sqrt{9}} + \frac{1}{\sqrt{10}}} = 1.2.$$
Javel. Dette betyr at $\hat{\mu}_x - \hat{\mu}_y$ er normalfordelt med forventning 0 og standardavvik 1.2. Vårt forsøk har jo gitt
$$\hat{\mu}_x - \hat{\mu}_y = 7.84 - 7.19 = 0.65,$$
som er $\frac{0.65}{1.2} = 0.55$ standardavvik.
Men hvis $\hat{\mu}_x - \hat{\mu}_y$ er normalfordelt som nevnt over, kan vi finne i tabellen at
$$P(-0.55\sigma < \hat{\mu}_x - \hat{\mu}_y < 0.55\sigma) = 0.4172.$$
Dette betyr at
$$P(\hat{\mu}_x - \hat{\mu}_y < -0.55\sigma \bigcup 0.55\sigma < \hat{\mu}_x - \hat{\mu}_y) = 0.5828.$$
Så det er altså hele 58.3% sannsynlighet for å få en målt forskjell på 0.65 kg eller mer på de to gruppene dersom de har samme gjennomsnittsvekt, og da er det ikke belegg for å forkaste denne hypotesen.

Ukjent σ

Hvis σ er ukjent, må vi estimere den. Da er det så enkelt som at vi må bruke formelen
$$\begin{aligned}\hat{\sigma} &= \sqrt{\frac{\sum_{n=1}^{n_1}(x_i - \hat{\mu}_y)^2 + \sum_{n=1}^{n_2}(y_i - \hat{\mu}_y)^2}{n_1 + n_2 - 2}} \\ &= \sqrt{\frac{\sum_{n=1}^{9}(x_i - 7.84)^2 + \sum_{n=1}^{10}(y_i - 7.19)^2}{9 + 10 - 2}} = 1.495,\end{aligned}$$
og så må vi bruke t-fordelingen med denne $\hat{\sigma}$. Vi har fremdeles at
$$\hat{\mu}_x - \hat{\mu}_y = 7.84 - 7.19 = 0.65,$$
som nå er $\frac{0.65}{1.495} = 0.435$ estimerte standardavvik. T-fordelingen med signifikans 10% gir oss at (slå opp på 9+10-2=17)
$$P(-1.333\sigma < \hat{\mu}_x - \hat{\mu}_y < 1.333\sigma) = 1 - 0.1 \cdot 2 = 0.80.$$
Nå er en målt forskjell på 0.435σ langt innenfor dette intervallet, og vi ser at også her er det ingen belegg for å forkaste hypotesen.