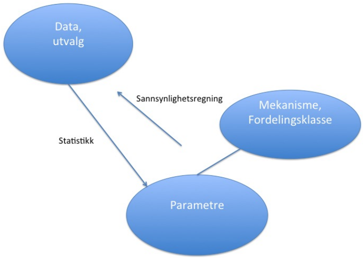
Statistikk

Jo Eidsvik

Matematiske fag, NTNU

Utvalg og Estimering

- Data er et utvalg: $X_1, X_2, \ldots, X_n$ uavhengige og identisk fordelt med tetthet eller punktsannsynlighet $f(x;\theta)$.
- Vi vil estimere modellparameter $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$.
- Vi antar typisk at fordelingstype (Normal, eksponensial, Poisson, eller lignende) er kjent fra erfaring omkring mekanismen.



Gjennomsnittet i et utvalg

$X_1, \ldots, X_n$ er uavhengige normalfordelte med $E(X_i) = \mu$, $Var(X_i) = \sigma^2$.

$$\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$$

1. $\bar{X}$ estimerer μ. Det er et **punktestimat**.
2. **Forventningsrett**: $E(\bar{X}) = \mu$.
3. **Varians går mot** 0: $Var(\bar{X}) = \sigma^2/n \to 0$ når $n \to \infty$.
4. $Z = \sqrt{n}\frac{\bar{X}-\mu}{\sigma}$, Z er normalfordelt med $E(Z) = 0$, $Var(Z) = 1$.

Vi kan bruke standard normalfordelingen til Z til å lage et **intervall** for estimering av μ, ikke bare et punktestimat.

Varians i et utvalg

$X_1, \dots, X_n$ er uavhengige normalfordelte med $E(X_i) = \mu$, $Var(X_i) = \sigma^2$.

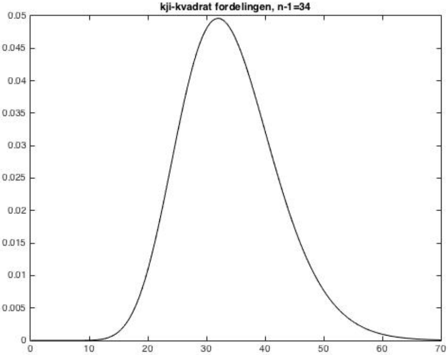
$$\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

1. S^2 estimerer σ^2.
2. **Forventningsrett**: $E(S^2) = \sigma^2$.
3. **Varians går mot** 0: $Var(S^2) \to 0$ når $n \to \infty$.
4. $Y = \frac{(n-1)S^2}{\sigma^2}$. Y er kji-kvadratfordelt med $n-1$ frihetsgrader.

Kji-kvadrat fordeling

$\chi^2_{34,0.05} = 22$, $\chi^2_{34,0.95} = 49$.



Eksempel - variasjon i bredde på akslinger

$X_1, \ldots, X_{35}$, Estimat for σ^2 er $S^2 = 0.0029$. Dette er bare et **punktestimat**.
Er det slik at $0.0029 \approx 0.005$? Eller er $0.0029 << 0.005$?
Vi kan bruke kji-kvadratfordelingen til Y til å lage et **intervall** for estimering av σ, ikke bare et punktestimat.

To-sample utvalg

Utvalg 1: $X_1, \ldots, X_{n_1}$ er uavhengige normalfordelte med $E(X_i) = \mu_1$, $Var(X_i) = \sigma^2$.
Utvalg 2: $Y_1, \ldots, Y_{n_2}$ er uavhengige normalfordelte med $E(Y_i) = \mu_2$, $Var(Y_i) = \sigma^2$.

Plasma ascorbic acid values

Nonsmokers		Smokers
0.97	1.16	0.48
0.72	0.86	0.71
1.00	0.85	0.98
0.81	0.58	0.68
0.62	0.57	1.18
1.32	0.64	1.36
1.24	0.98	0.78
0.99	1.09	1.64
0.90	0.92	
0.74	0.78	
0.88	1.24	
0.94	1.18	

To-sample utvalg estimering

Estimator av differansen $\mu_1 - \mu_2$.

$$\bar{X} - \bar{Y} = \frac{1}{n_1}(X_1 + \ldots + X_{n_1}) - \frac{1}{n_2}(Y_1 + \ldots + Y_{n_2})$$

1. $\bar{X} - \bar{Y}$ er et **punktestimat** for differansen.
2. **Forventningsrett**: $E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$.
3. **Varians går mot** 0: $Var(\bar{X} - \bar{Y}) = \sigma^2(1/n_1 + 1/n_2) \to 0$ når $n_1 \to \infty$ og $n_2 \to \infty$.
4. $Z = \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{1/n_1+1/n_2}}$, er normalfordelt med forventning 0 og varians 1.

To-sample utvalg estimering : askorbinsyre hos røykere / ikke røykere

Estimator av differansen $\mu_1 - \mu_2$.

$$\bar{X} - \bar{Y} = \frac{1}{n_1}(X_1 + \ldots + X_{n_1}) - \frac{1}{n_2}(Y_1 + \ldots + Y_{n_2})$$

1. $\bar{X} - \bar{Y}$ er et **punktestimat** for differansen.
2. **Forventningsrett**: $E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$.
3. **Varians går mot** 0: $Var(\bar{X} - \bar{Y}) = \sigma^2(1/n_1 + 1/n_2) \to 0$ når $n_1 \to \infty$ og $n_2 \to \infty$.
4. $Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}}$, er normalfordelt med forventning 0 og varians 1.

Her er $n_1 = 24$, $n_2 = 8$, $\bar{x} = 0.98$, $\bar{y} = 0.92$, $\sigma = 0.26$. Hva om vi antar $\mu_1 - \mu_2 = 0$?

$$z = \frac{0.98 - 0.92}{0.26\sqrt{1/24 + 1/8}} = 0.56$$

Det er ikke veldig stort tall i en standard normalfordeling!

To-sample utvalg estimering

Estimator av differansen $\mu_1 - \mu_2$.

$$\bar{X} - \bar{Y} = \frac{1}{n_1}(X_1 + \ldots + X_{n_1}) - \frac{1}{n_2}(Y_1 + \ldots + Y_{n_2})$$

1. $\bar{X} - \bar{Y}$ er et **punktestimat** for differansen.
2. **Forventningsrett**: $E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$.
3. **Varians går mot** 0: $Var(\bar{X} - \bar{Y}) = \sigma^2(1/n_1 + 1/n_2) \rightarrow 0$ når $n_1 \rightarrow \infty$ og $n_2 \rightarrow \infty$.
4. $Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}}$, er normalfordelt med forventning 0 og varians 1.

Vi kan bruke standard normalfordelingen til Z til å lage et **intervall** for estimering av $\mu_1 - \mu_2$, ikke bare et punktestimat.