

TMA4240 Statistikk H2016 [19]

Prediksjonsintervall [9.6]

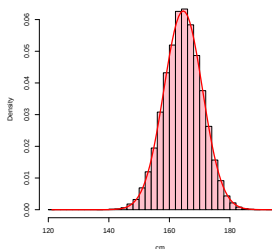
To utvalg: estimere differanse mellom to forventningsverdier [8.4, 9.8]

Mette Langaas

Institutt for matematiske fag, NTNU

wiki.math.ntnu.no/emner/tma4240/2016h/start/

Hvor høye er kvinnelige studenter ved NTNU?



- ▶ Populasjon: alle kvinnelige studenter ved NTNU.
- ▶ Tilfeldig utvalg: kvinner i TMA4240 som har svart på spørreundersøkelse ved oppstart av kurset.
 - ▶ $n=149$
 - ▶ Utvalgsgjennomsnitt: 169.5 cm
 - ▶ Utvalgsstandardavvik: 6.6 cm
- ▶ Sist: 95% konfidensintervall for forventet høyde: $[168.4, 170.6]$. Vi har stor tiltro (95% tillit) til at sann forventet høyde ligger i dette intervallet.

Konfidensintervall for μ

$X_1, X_2, \dots, X_n$ u.i.f. $N(\mu, \sigma^2)$. $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ og $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$.

σ kjent:

$$\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

σ ukjent:

$$\bar{x} - t_{\frac{\alpha}{2}, (n-1)} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\frac{\alpha}{2}, (n-1)} \frac{s}{\sqrt{n}}$$

Hvis vi ønsker konfidensintervall for θ , finner vi to observatorer L og U slik at

$$P(L < \theta < U) = 1 - \alpha \text{ for } 0 < \alpha < 1$$

og vi har sannsynlighet $1 - \alpha$ for å lage et intervall som dekker den sanne θ . Intervallet laget fra et utvalg, $[\ell, u]$ kaller vi et $100(1 - \alpha)\%$ konfidensintervall for θ .

Kritiske verdier i standard normalfordelingen

$$P(Z > z_{\alpha}) = \alpha$$

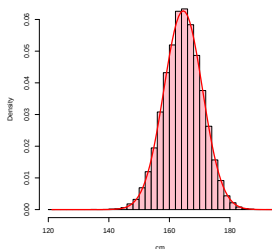
α	z_{α}
.2	0.842
.15	1.036
.1	1.282
.075	1.440
.05	1.645
.04	1.751
.03	1.881
.025	1.960
.02	2.054
.01	2.326
.005	2.576
.001	3.090
.0005	3.291
.0001	3.719
.00005	3.891
.00001	4.265
.000005	4.417
.000001	4.753

Kritiske verdier i t -fordelingen

$$P(T > t_{\alpha, \nu}) = \alpha$$

$\nu \backslash \alpha$	.150	.100	.075	.050	.025	.010	.005	.001	.0005
1	1.963	3.078	4.165	6.314	12.706	31.821	63.657	318.309	636.619
2	1.386	1.886	2.282	2.920	4.303	6.965	9.925	22.327	31.599
3	1.250	1.638	1.924	2.353	3.182	4.541	5.841	10.215	12.924
4	1.190	1.533	1.778	2.132	2.776	3.747	4.604	7.173	8.610
5	1.156	1.476	1.699	2.015	2.571	3.365	4.032	5.893	6.869
6	1.134	1.440	1.650	1.943	2.447	3.143	3.707	5.208	5.959
7	1.119	1.415	1.617	1.895	2.365	2.998	3.499	4.785	5.408
8	1.108	1.397	1.592	1.860	2.306	2.896	3.355	4.501	5.041
9	1.100	1.383	1.574	1.833	2.262	2.821	3.250	4.297	4.781
10	1.093	1.372	1.559	1.812	2.228	2.764	3.169	4.144	4.587
11	1.088	1.363	1.548	1.796	2.201	2.718	3.106	4.025	4.437
12	1.083	1.356	1.538	1.782	2.179	2.681	3.055	3.930	4.318
13	1.079	1.350	1.530	1.771	2.160	2.650	3.012	3.852	4.221
14	1.076	1.345	1.523	1.761	2.145	2.624	2.977	3.787	4.140
15	1.074	1.341	1.517	1.753	2.131	2.602	2.947	3.733	4.073
16	1.071	1.337	1.512	1.746	2.120	2.583	2.921	3.686	4.015
17	1.069	1.333	1.508	1.740	2.110	2.567	2.898	3.646	3.965
18	1.067	1.330	1.504	1.734	2.101	2.552	2.878	3.610	3.922
19	1.066	1.328	1.500	1.729	2.093	2.539	2.861	3.579	3.883
20	1.064	1.325	1.497	1.725	2.086	2.528	2.845	3.552	3.850
21	1.063	1.323	1.494	1.721	2.080	2.518	2.831	3.527	3.819
22	1.061	1.321	1.492	1.717	2.074	2.508	2.819	3.505	3.792
23	1.060	1.319	1.489	1.714	2.069	2.500	2.807	3.485	3.768
24	1.059	1.318	1.487	1.711	2.064	2.492	2.797	3.467	3.745
25	1.058	1.316	1.485	1.708	2.060	2.485	2.787	3.450	3.725
26	1.058	1.315	1.483	1.706	2.056	2.479	2.779	3.435	3.707
27	1.057	1.314	1.482	1.703	2.052	2.473	2.771	3.421	3.690
28	1.056	1.313	1.480	1.701	2.048	2.467	2.763	3.408	3.674
29	1.055	1.311	1.479	1.699	2.045	2.462	2.756	3.396	3.659
30	1.055	1.310	1.477	1.697	2.042	2.457	2.750	3.385	3.646
35	1.052	1.306	1.472	1.690	2.030	2.438	2.724	3.340	3.591
40	1.050	1.303	1.468	1.684	2.021	2.423	2.704	3.307	3.551
50	1.047	1.299	1.462	1.676	2.009	2.403	2.678	3.261	3.496
60	1.045	1.296	1.458	1.671	2.000	2.390	2.660	3.232	3.460
80	1.043	1.292	1.453	1.664	1.990	2.374	2.639	3.195	3.416
100	1.042	1.290	1.451	1.660	1.984	2.364	2.626	3.174	3.390
120	1.041	1.289	1.449	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.036	1.282	1.440	1.645	1.960	2.326	2.576	3.090	3.291

Hvor høye er kvinnelige studenter ved NTNU?



- ▶ Populasjon: alle kvinnelige studenter ved NTNU.
- ▶ Tilfeldig utvalg: kvinner i TMA4240 som har svart på spørreundersøkelse ved oppstart av kurset.
 - ▶ $n=149$
 - ▶ Utvalgsgjennomsnitt: 169.5 cm
 - ▶ Utvalgsstandardavvik: 6.6 cm
- ▶ Sist: 95% konfidensintervall for forventet høyde: [168.4,170.6]. Vi har stor tiltro (95% tillit) til at sann forventet høyde ligger i dette intervallet.
- ▶ Nå: Ønsker intervall der det er 95% sannsynlighet for *ny observasjon* vil ligge.

Prediksjonsintervall for fremtidig observasjon, normalfordeling

- For en normalfordeling med ukjent forventningsverdi μ , og ukjent varians σ^2 , er et $(1-\alpha)100\%$ prediksjonsintervall for en fremtidig observasjon x_0 gitt som

$$\bar{x} - t_{\frac{\alpha}{2},(n-1)}s\sqrt{1 + \frac{1}{n}} < x_0 < \bar{x} + t_{\frac{\alpha}{2},(n-1)}s\sqrt{1 + \frac{1}{n}}$$

hvor $t_{\frac{\alpha}{2},(n-1)}$ er verdien i t-fordelingen med $n - 1$ frihetsgrader som har areal $\frac{\alpha}{2}$ til høyre, dvs. $P(T > t_{\frac{\alpha}{2},(n-1)}) = \frac{\alpha}{2}$, og $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$

Prediksjonsintervall for fremtidig observasjon, normalfordeling

- For en normalfordeling med ukjent forventningsverdi μ , men kjent varians σ^2 , er et $(1-\alpha)100\%$ prediksjonsintervall for en fremtidig observasjon x_0 gitt som

$$\bar{x} - z_{\frac{\alpha}{2}}\sigma\sqrt{1 + \frac{1}{n}} < x_0 < \bar{x} + z_{\frac{\alpha}{2}}\sigma\sqrt{1 + \frac{1}{n}}$$

hvor $z_{\frac{\alpha}{2}}$ er verdien i normal-fordelingen som har areal $\frac{\alpha}{2}$ til høyre, dvs. $P(Z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$.

Prediksjonsintervall - generelt

Hvis vi ønsker prediksjonsintervall for en ny observasjon X_0 , finner vi to observatorer L og U (basert på vårt sample) slik at

$$P(L < X_0 < U) = 1 - \alpha \text{ for } 0 < \alpha < 1$$

og vi har sannsynlighet $1 - \alpha$ for å lage et intervall som dekker en ny observasjon X_0 . Intervallet laget fra et utvalg, $[\ell, u]$ kaller vi et $100(1 - \alpha)\%$ prediksjonsintervall for X_0 .

Hva gjenstår innen estimering?

- ▶ Sammenligne bensinforbruk for to biltyper - parameterestimat og konfidensintervall.
 - ▶ To uavhengige utvalg: differanse i forventningsverdi, normalfordelte populasjoner. i) Kjente varianser, ii) ukjente varianser.
- ▶ Lag et konfidensintervall for differansen i populasjonsgjennomsnittet i kroppsfettprosent etter pedikyr og før pedikyr.
 - ▶ Parutvalg (to avhengige utvalg).
- ▶ Lag et konfidensintervall for andelen studenter som synes de er flinkere å kjøre bil enn landsgjennomsnittet.
 - ▶ Ett utvalg (binomisk): andel
- ▶ Synes flere unge menn enn kvinner at de er flinkere enn landsgjennomsnittet til å kjøre bil?
 - ▶ To utvalg (binomisk): andeler
- ▶ Kraften som trengs til å åpne en vinkork
 - ▶ Ett utvalg: undersøke varians.

To utvalg: statistisk situasjon

- ▶ Ønsker å sammenligne to populasjoner basert på et u.i.f. utvalg fra hver populasjon.
- ▶ Studerer en egenskap som kan sies å være normalfordelt i hver populasjon,
- ▶ og ønsker å anslå differansen mellom forventningsverdien i de to populasjonene
- ▶ og et intervall der vi har stor tillit til at den sanne differansen i forventningsverdiene ligger.
- ▶ Sammenligningene kan være parvise eller ikke parvise.

To uavhengige utvalg: Bensinforbruk

Problemstilling:

- ▶ Vil sammenligne to biltyper, 1 og 2, mhp bensinforbruk.
- ▶ Utvalg 1: X_{1i} : km/liter for biltype nummer i , type 1.
- ▶ Utvalg 2: X_{2j} : km/liter for biltype nummer j , type 2.
- ▶ De to utvalgene er uavhengige.
- ▶ Anta at X_{1i} er normalfordelt med ukjent μ_1 og kjent $\sigma_1 = \sqrt{2}$ km/liter.
- ▶ Anta at X_{2j} er normalfordelt med ukjent μ_2 og kjent $\sigma_2 = \sqrt{3}$ km/liter.

Observasjoner:

- ▶ $n_1 = 12$ målinger på biltype 1, med gjennomsnitt $\bar{x}_1 = 10$ km/liter.
- ▶ $n_2 = 10$ målinger på biltype 2, med gjennomsnitt $\bar{x}_2 = 8$ km/liter.

To utvalg: estimatorer

- ▶ $X_1^1, X_2^1, \dots, X_{n_1}^1$ er et tilfeldig utvalg fra en populasjon som beskrives av en normalfordeling med forventning μ_1 og varians σ_1^2 .
- ▶ $X_1^2, X_2^2, \dots, X_{n_2}^2$ er et tilfeldig utvalg fra en populasjon som beskrives av en normalfordeling med forventning μ_2 og varians σ_2^2 .
- ▶ Estimator for $\mu_1 - \mu_2$:
 $\hat{\mu}_1 - \hat{\mu}_2 = \bar{X}_1 - \bar{X}_2 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i^1 - \frac{1}{n_2} \sum_{j=1}^{n_2} X_j^2$ (intuitiv og SME).
- ▶ $\bar{X}_1 - \bar{X}_2$ er normalfordelt med

$$\begin{aligned} E(\bar{X}_1 - \bar{X}_2) &= \mu_1 - \mu_2 \\ \text{Var}(\bar{X}_1 - \bar{X}_2) &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \end{aligned}$$

- ▶ Hvis σ_1 og σ_2 er kjente så er Z standard normalfordelt.

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

To uavhengige utvalg: differanse av to gjennomsnitt

Sentralgrenseteoremet:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

går mot standard normalfordelingen, $n(z; 0, 1)$, når n_1 og $n_2 \rightarrow \infty$.

To utvalg: konfidensintervall for $\mu_1 - \mu_2$ når σ_1^2 og σ_2^2 er kjente

- ▶ Hvis $\bar{X}_1$ og $\bar{X}_2$ er gjennomsnittene til to tilfeldig utvalg av størrelse n_1 og n_2 fra populasjoner med kjent varians σ_1^2 og σ_2^2 , så er et $(1-\alpha)100\%$ konfidensintervall for $\mu_1 - \mu_2$

$$\begin{aligned} (\bar{x}_1 - \bar{x}_2) - z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} &< (\mu_1 - \mu_2) < \\ (\bar{x}_1 - \bar{x}_2) + z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \end{aligned}$$

hvor $z_{\frac{\alpha}{2}}$ er verdien i standard normalfordelingen som har areal $\frac{\alpha}{2}$ til høyre, dvs. $P(Z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$.

To utvalg: Bensinforbruk

95% konfidensintervall for $\mu_1 - \mu_2$.

- ▶ Punktestimator: $\hat{\mu}_1 - \hat{\mu}_2 = \bar{X}_1 - \bar{X}_2$.
- ▶ Punktestimat: $\bar{x}_1 - \bar{x}_2 = 2$ km/liter.
- ▶ 95% konfidensintervall: $\alpha = 0.05$, $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$,

$$\left[2 - 1.96\sqrt{\frac{2}{12} + \frac{3}{10}}, 2 + 1.96\sqrt{\frac{2}{12} + \frac{3}{10}}\right] = [0.66, 3.34]$$

To utvalg: σ_1^2 og σ_2^2 ukjente

- ▶ Estimatorer for σ_1^2 og σ_2^2 :

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^2 \quad \text{og} \quad S_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (X_{2j} - \bar{X}_2)^2$$

- ▶ $(1 - \alpha)100\%$ konfidensintervall for $\mu_1 - \mu_2$:

$$[(\bar{X}_1 - \bar{X}_2) \pm t_{\frac{\alpha}{2}, \nu} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}]$$

der

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{[(s_1^2/n_1)^2/(n_1 - 1)] + [(s_2^2/n_2)^2/(n_2 - 1)]}$$

- ▶ Dette er et tilnærmet resultat, og vil bli oppgitt enten som formel eller som numerisk verdi hvis det blir gitt i en eksamensoppgave.

To utvalg: $\sigma_1^2 = \sigma_2^2$, men ukjente

Tidligere i pensum - men nå ikke pensum. Hvorfor?

Fordi denne situasjonen veldig sjelden brukes i praksis. Man vet ofte ikke at de to variansene er like. Grunnen til at dette har vært med tidligere er at ved å lage en felles estimator for σ^2 som kalles S_{pooled} som har en klar sammenheng med en khikvadratfordeling med $n_1 + n_2 - 2$ frihetsgrader kan man finne en eksakt t-fordelt størrelse med $n_1 + n_2 - 2$ frihetsgrader (ikke bare tilnærmet slik vi fant).

Men, dette gav alt for mye fokus på khikvadratfordeling og blir ikke så mye brukt nå som vi har software som regner ut de tilnærmede frihetsgradene vi har når de to variansene er ulike.

Differanse mellom høyde til menn og kvinner

Oppgave: Gi et estimat og 95% konfidensintervall for differansen til forventet høyde til mannlige og kvinnelige studenter ved NTNU.

Populasjon 1 (menn): $n_1 = 216$, $\bar{x}_1 = 183.8$, $s_1 = 6.9$

Populasjon 2 (kvinner): $n_2 = 149$, $\bar{x}_2 = 169.5$, $s_2 = 6.6$

- ▶ Punktestimator: $\hat{\mu}_1 - \hat{\mu}_2 = \bar{X}_1 - \bar{X}_2$.
- ▶ Punktestimat: $\bar{x}_1 - \bar{x}_2 = 183.8 - 169.5 = 14.3$ cm.
- ▶ Frihetsgrader for Student-t:
$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{[(s_1^2/n_1)/(n_1-1)] + [(s_2^2/n_2)/(n_2-1)]} = 328.04$$
 som rundes ned til 328 (men 328 finnes jo ikke i tabellen).
- ▶ 95% konfidensintervall: $\alpha = 0.05$, $t_{\frac{\alpha}{2}, \nu} = t_{0.025, 328} = 1.967$ (men dette må dere finne i matlab, det kan dere ikke finne i tabellen - må dere bruke tabellen bør dere velge 1.96),

$$[14.3 \pm 1.967 \sqrt{\frac{6.9^2}{216} + \frac{6.6^2}{149}} = [12.94, 15.76]$$

To uavhengige utvalg: oppsummering konfidensintervall

$(1 - \alpha)100\%$ konfidensintervall for $\mu_1 - \mu_2$:

- ▶ når σ_1^2 og σ_2^2 er kjent:

$$[(\bar{x}_1 - \bar{x}_2) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}]$$

- ▶ når σ_1^2 og σ_2^2 er ukjente:

$$[(\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2}, \nu} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}]$$

der

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{[(s_1^2/n_1)^2/(n_1 - 1)] + [(s_2^2/n_2)^2/(n_2 - 1)]}$$

To utvalg: eksempler

- ▶ Betong: to ulike oppskrifter, A og B, skal sammenlignes. Hvor stor forskjell er det i styrken (“crushing strength”) for betong fra oppskrift A og fra oppskrift B?
- ▶ Algoritmer: to ulike sorteringsalgoritmer skal sammenlignes. Hvilken algoritme har kortest kjøretid for en spesiell problemstilling?
- ▶ Sykdom: tester ut ny blodtrykksmedisin. Hvor mye bedre er den enn nåværende markedsledende blodtrykksmedisin?
- ▶ Kosthold: hvor stor vektreduksjon vil man oppleve ved å følge diett A i et halvt år?
- ▶ Bildekk: to typer dekk, A og B, skal sammenlignes mhp slitasje. Kan enten sette både A og B-dekk på hver bil eller noen biler med A og noen biler med B.