Sammenligning av to grupper : Two-sample problems

∞ populasjon med to grupper

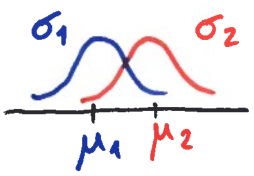
gruppe 1: Fordeling 1 med parameter μ_1 og σ_1

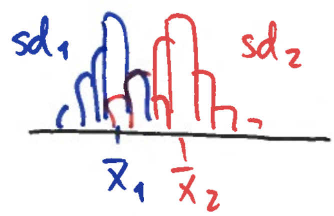
Gruppe 2: Fordeling 2 med parameter μ_2 og σ_2

$\longrightarrow$ utvalg n uavh. data

n_1 data fra gruppe 1, $\bar{x}_1$, sd_1

n_2 data fra gruppe 2, $\bar{x}_2$, sd_2





(A)

Da vil

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1) \quad \text{hvis } \sigma \text{ kjent}$$

Under $H_0 : \mu_1 = \mu_2$, er

$$z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

z testobservator
z-test

(B)

Eller

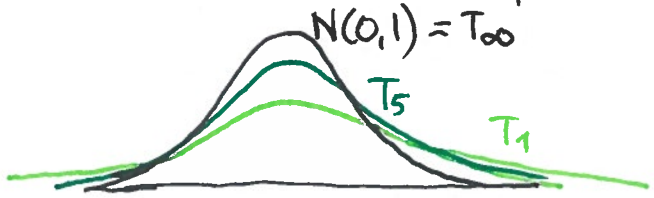
$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{sd_1^2}{n_1} + \frac{sd_2^2}{n_2}}} \sim T_{\min(n_1-1,\, n_2-1)}$$

Under $H_0 : \mu_1 = \mu_2$, er

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{sd_1^2}{n_1} + \frac{sd_2^2}{n_2}}} \sim T_{\min(n_1-1,\, n_2-1)}$$

t observator
t-test

Husk at jo større n blir, jo mer nærmer T-fordelingen seg N(0,1)



(c) Eller, hvis vi ikke vet σ, men kan anta at σ_1 er lik σ_2, kan vi erstatte nevneren i uttrykket, og få at

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - \overbrace{(\mu_1 - \mu_2)}^{=0 \text{ under } H_0}}{S_{pooled} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim T_{n_1+n_2-2}$$

t-observator
t-test

$$S^2_{pooled} = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

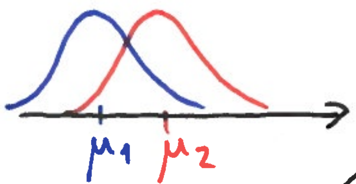
7.64 Et 95% konfidensintervall (KI) for forskjellen på to forventningsverdier er [0.8, 2.3]. Hva kan du si om hypotesetesten for $H_0: \mu_1 = \mu_2$ mot $H_A: \mu_1 \neq \mu_2$?

Denne situasjonen er basert på følgende:

∞

I populasjonen

Her er det to grupper, gruppe 1 og gruppe 2, som har fordelinger som ser omtrent sånn ut:



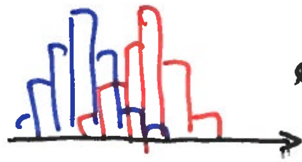
(for eksempel M & K)

for eksempel
hvilepuls
Blodtrykk ...

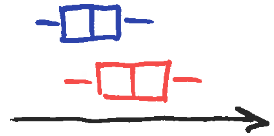
trekker vi et utvalg fra populasjonen (noen fra hver av de to gruppene):

utvalg: kjent.
= observasjoner.

Nå kan vi undersøke den observerte fordelingen i de to gruppene:



eller



Disse gruppene oppsummeres godt med

μ_1 = forventningsverdien "populasjonsgjennomsnittet" i gruppe 1 og

μ_2 = forventningsverdien "populasjonsgjennomsnittet" i gruppe 2

Men begge er ukjente, og må estimeres. Vi vet heller ikke hva forskjellen på de to gruppene er, eller om det i det hele tatt er noen forskjell.

Derfor

Disse to gruppene oppsummeres godt med

$\bar{X}_1$ = gjennomsnittet i gruppe 1 og

$\bar{X}_2$ = gjennomsnittet i gruppe 2.

Vi bruker $\bar{X}_1$ og $\bar{X}_2$ som estimater for μ_1 og μ_2. Hvis $\bar{X}_1$ og $\bar{X}_2$ er ganske like, tror vi at H_0 er riktig, men hvis $\bar{X}_1$ og $\bar{X}_2$ er ganske ulike, stoler vi ikke lenger på H_0. Hvor forskjellige må $\bar{X}_1$ og $\bar{X}_2$ være for at vi skal forkaste H_0?

Legg merke til at

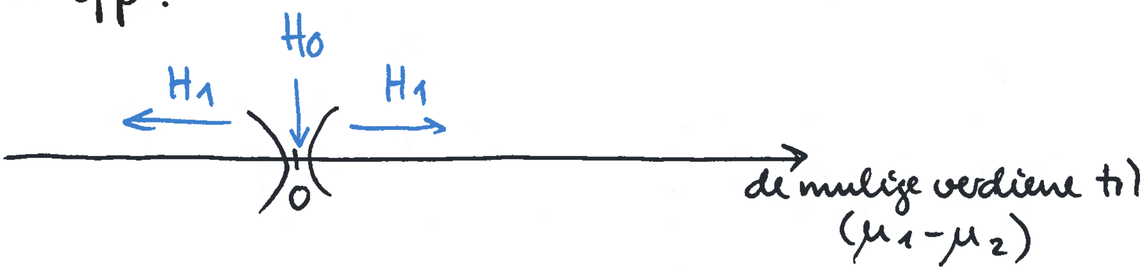
$H_0 : \mu_1 = \mu_2$ kan skrives som

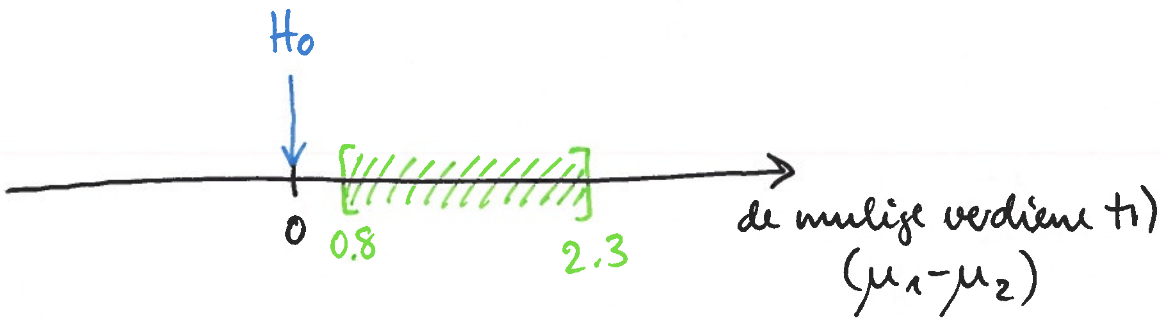
$H_0 : (\mu_1 - \mu_2) = 0$, og dermed $H_A : (\mu_1 - \mu_2) \neq 0$

Mao: Hvor stor må differansen mellom $\bar{X}_1$ og $\bar{X}_2$ være for at vi skal forkaste H_0?

Intervallet [0.8, 2.3] har 1.55 i midten, og en konfidensgrad på 95%. Vi er altså 95% sikre på at intervallet inneholder den sanne differansen mellom μ_1 og μ_2.

Tegner det opp:





H_0-verdien $(\mu_1 - \mu_2) = 0$ er ikke i intervallet. Forkast H_0.

Dette tilsvarer en tosidig hypotesetest med $\alpha = 0.05$

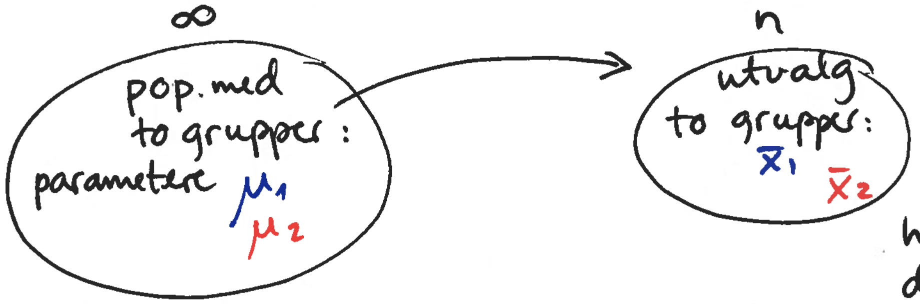
b) Med et større utvalg vil feilmarginene ("slingringsmonnet") bli mindre, og konfidensintervallet blir smalere.

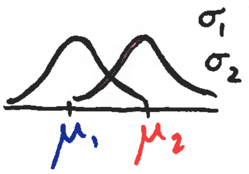
<u>7.65</u> a) En signifikanstest for å sammenligne to forventningsverdier,

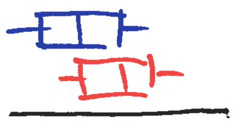
$$H_0 : \mu_1 = \mu_2 \iff (\mu_1 - \mu_2) = 0$$
$$H_A : \mu_1 \neq \mu_2 \iff (\mu_1 - \mu_2) = 0$$

ga en t-observator-verdi på $t_{10} = -1.97$.

Hva ligger bak? Igjen,







hvis n liten
data ~ Normalf.
σ_1, σ_2 ukjent:
sd_1, sd_2 estimater
for σ_1 og σ_2

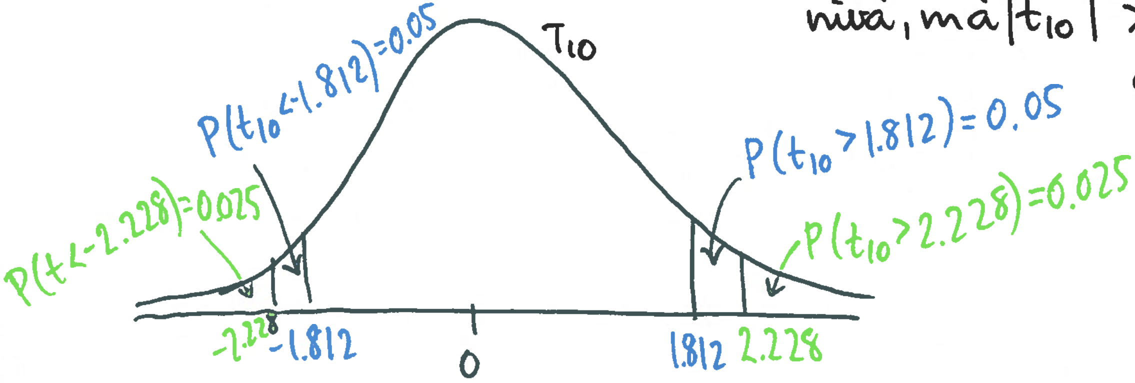
Da vil
$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{sd_1}{n_1} + \frac{sd_2}{n_2}}} \sim T_{n_1 - 1}$$

Når H_0 er sann, er $\mu_1 - \mu_2 = 0$, og da er

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{sd_1}{n_1} + \frac{sd_2}{n_2}}} \sim T_{n_1 - 1}$$

Vi får oppgitt at $t_{10} = -1.97$, og må sammenligne det med tall i Table D: For å forkaste H_0 på $\alpha = 0.05$-nivå, må $|t_{10}| > 2.228$, og det er den ikke.

Behold H_0



7.68 $n = 1839$ studenter

Studietid per uke : $\bar{x}_{st} = 706$ min, $sd_{st} = 526$ min

Facebook-tid per dag : $\bar{x}_{fb} = 106$ min, $sd_{fb} = 93$ min

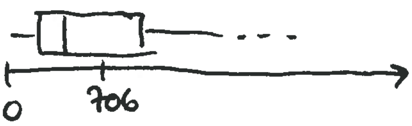
8% : 0 tid på fb

Begge $\bar{x}$ og sd viser at de individuelle dataene er svært skjevfordelte.

Noterer : n er stor.

a) μ_{st} skal estimeres.

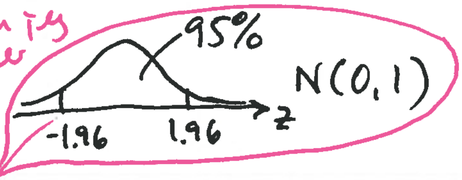
Individuelle data ser kanskje noe sånt ut :



Selv om $\bar{x}_{st}$ ikke er en god oppsummering av det typiske i disse dataene, er n stor nok til at $\bar{x}_{st}$ ligner på μ_{st} og at

$$z = \frac{\bar{x}_{st} - \mu_{st}}{sd_{st}/\sqrt{n}} \overset{\text{tiln}}{\sim} N(0,1)$$

Her er n så stor at jeg går rett på N(0,1), selv om jeg estimerer σ_{st} med sd_{st}



Igjen : $P(-1.96 < z < 1.96) = 0.95$ i en N(0,1) fordeling

Altså :

$$P\left(-1.96 < \frac{\bar{x} - \mu}{sd/\sqrt{n}} < 1.96\right) \approx 0.95$$

$$\vdots$$

$$P\left(\bar{x} - 1.96\frac{sd}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{sd}{\sqrt{n}}\right) \approx 0.95$$

95% KI for μ_{st} : $706 \pm 1.96 \cdot \frac{526}{\sqrt{1839}} \rightarrow [682, 730]$

b) Fb-tid per uke : $\bar{x}_f = 742$, $sd_f = 651$ (antar at de 8% er innbakt i $\bar{x}_{fb}$)

Samme vurderinger om skjevhet & n

95% KI for μ_f, forventet fb-tid per uke :

$742 \pm 1.96 \cdot \frac{651}{\sqrt{1839}} \longrightarrow [712, 772]$ (huff.)

c) Nei, n er stor.

7.70 Ønsker å sammenligne $\bar{x}_{st} = 706$ med $\bar{x}_f = 742$

a) Estimat for forskjellen: $\bar{x}_f - \bar{x}_{st} = 742 - 706 = \underline{36}$

En halvtime mer på fb enn på studier per uke.

b) Kan ikke bruke t-test (eller z-test) for å sml, fordi disse dataene ikke er uavhengige.

Både z-test og t-test er basert på uavh. data, både innad og mellom gruppene.

7.75 To grupper: Late meal og early meal. Uavhengige deltakere.

$n = 200$ (late meal) $n = 202$ (early meal)

a) Kan vi bruke t-test her?

Ja, selv om den deskriptive statistikken tyder på at data er litt skjeve, er n ganske ok. σ må estimeres med sd, og da er t-test naturlig.

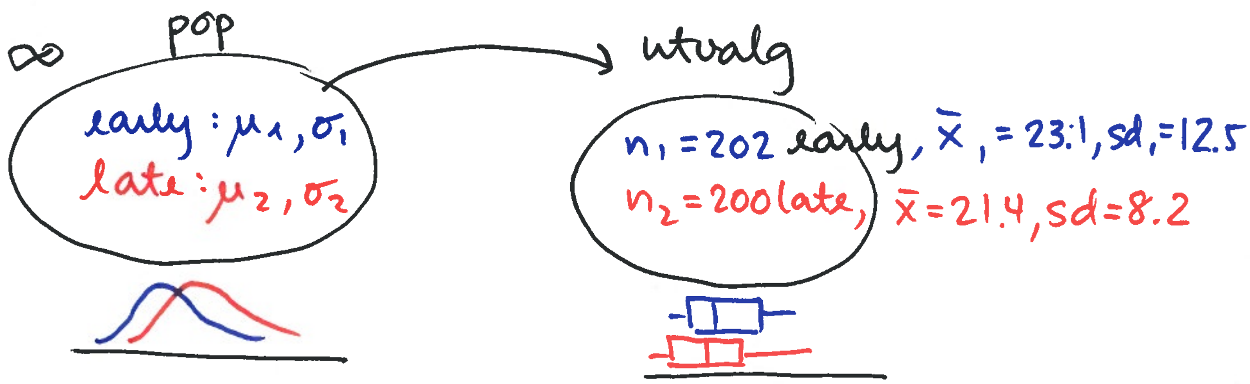
b) μ_1 er forventet fettinntak blant tidlig-spisere

μ_2 er forventet fettinntak blant sent-spisere

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

c) Signifikansnivå 0.05



t-observator:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{sd_1^2}{n_1} + \frac{sd_2^2}{n_2}}} = \frac{23.1 - 21.4}{\sqrt{\frac{12.5^2}{202} + \frac{8.2^2}{200}}} = 1.614$$

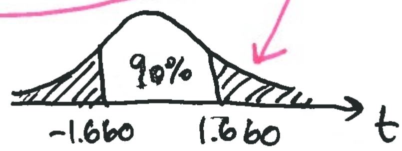
Må sammenlignes med T_{199} nesten $N(0,1)$

Finner bare T_{100} og T_{1000} i Table D

199 er nærmest 100

$$p\text{-verdi} = P\left(\begin{matrix}\text{minst like ekstreme observasjoner} \\ \text{som det vi har observert}\end{matrix} \middle| H_0\right)$$

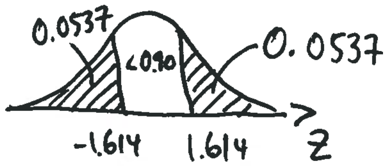
$$= P(|t| > 1.614) \rightarrow C \text{ litt mindre enn } 90\%$$



når n er såpass stor, ligner T på N(0,1)

$$\text{jfr } P(|z| > 1.614) = 2 \cdot 0.0537 = \underline{0.1074}$$

$$= P(z < -1.614) + (1 - P(z < 1.614)) = 0.0537 + (1 - 0.9463)$$



Dette er ikke signifikant, og H_0 beholdes.

$C < 0.9$ betyr at konfidensgraden er for lav til å aksepteres når vi har valgt $\alpha = 0.05$, altså konfidensgrad 0.95.

Estimatet vi får for p-verdien, basert på at n er stor, og at $T_{198} \approx N(0,1)$, er 0.11, og det er mye større enn 0.05.

d) 95% KI for $(\mu_1 - \mu_2)$, altså den forventede forskjellen i fettinntak: Bruker igjen at $T_{198} \approx N(0,1)$. Da er

$$P\left(-1.96 < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{sd_1^2}{n_1} + \frac{sd_2^2}{n_2}}} < 1.96\right) \approx 0.95$$

$\vdots$

95% KI for $(\mu_1 - \mu_2)$:

$$(\bar{X}_1 - \bar{X}_2) \pm 1.96 \cdot \sqrt{\frac{sd_1^2}{n_1} + \frac{sd_2^2}{n_2}} \longrightarrow 1.7 \pm 1.96 \cdot 1.05$$

$$\underline{[-0.36, 3.76]}$$

Dette intervallet inneholder også H_0-verdien 0, så det kan også brukes til å konkludere med <u>Behold H_0</u>