Ny og bedre versjon 2020

MAT110

Statistikk 1

Kompendium 2020, del 3

Per Kristian Rekdal
Bård-Inge Pettersen

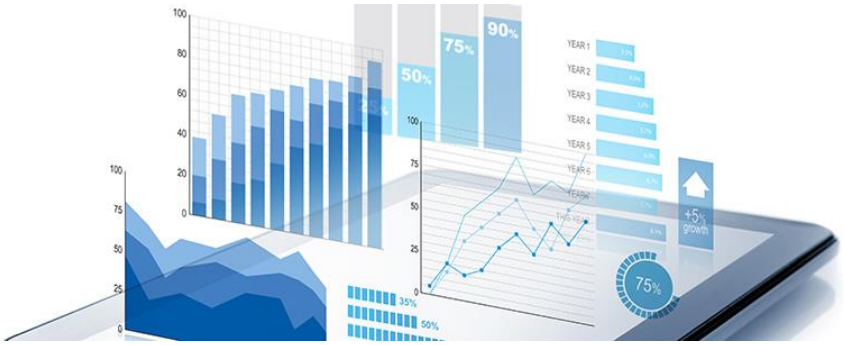
Høgskolen i Molde
Vitenskapelig høgskole i logistikk

Dette kompendiet er tillatt å ha med på eksamen.

Innhold

Kapittel 6

Estimering og konfidensintervaller



Figur 6.1: Estimering og konfidensintervaller.

6.1 Motivasjon - statistisk inferens

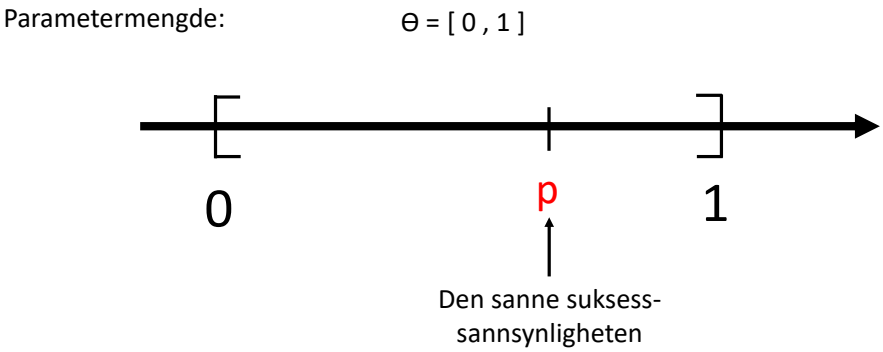
Vi har gjennomført en forsøksrekke på et tilfeldig utvalg av $n = 100$ pasienter som har fått et nytt legemiddel. Utfallet av et forsøk var enten suksess dersom effekten ble målt til 0.85 eller høyere og fiasko hvis ikke.

Statistisk modell:

Den statistiske modellen for de stokastiske forsøksvariablene $X_1, X_2, \ldots, X_{100}$ og populasjonsvariabelen X er gitt ved: (se lign.(5.64))

$$X_1,\ X_2,\ X_3,\ \ldots\ ,X_n \overset{\text{i.i.d.}}{\sim} X \sim \text{Ber}[p]_{p\in[0,1]} \tag{6.1}$$

hvor $p \in \Theta = [0, 1]$ er parametermengden og $V = \{0, 1\}$ verdimengden for modellen.

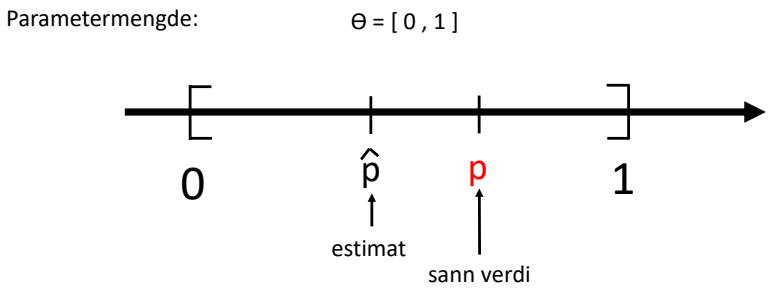


Figur 6.2: Mengden $\Theta = [0, 1]$.

To viktige størrelser:

$$p \quad = \quad \text{den sanne (og ukjente) suksess-sannsynligheten for at legemiddel gir effekt} \geq 0.85 \tag{6.2}$$

$$\hat{p} \quad = \quad \text{et estimat for den sanne suksess-sannsynligheten } p \tag{6.3}$$

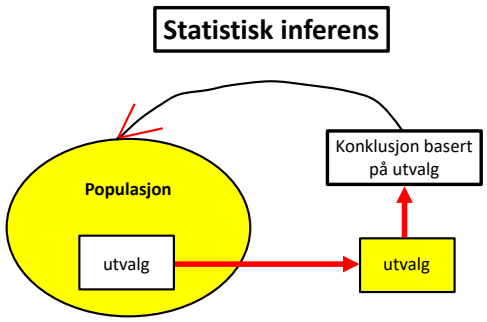


Figur 6.3: p VS $\hat{p}$.

Vi er nå klare for å gjennomføre

statistisk inferens

dvs. vi er klare for å trekke konklusjoner (=inferere) basert på forsøksrekken om legemiddelets for hele populasjonen - altså for *enhver* pasient med den aktuelle lidelsen.



Figur 6.4: Statistisk inferens.

1) Første tilnærming - **inferens**:

Det viktigste målet er å finne et *estimat* $\hat{p}$ for den sanne suksess-sannsynligheten p.

Fra tabell 5.3 på side 272, dvs. resultatet om de ble friske eller ikke med $\overbrace{\text{0'ere og 1'ere}}^{\text{tabell 5.3}}$, finner man at det er 88 av 100 pasienter som ble friske. *Relativfrekvensen* er dermed:

$$\hat{p} = f_r = \frac{\text{gunstige}}{\text{mulige}} = \frac{\text{antall som ble friske}}{\text{antall som fikk medisin}} = \frac{88}{100} = 0.88 \qquad (6.4)$$

siden 88 av de 100 testede $\underbrace{\text{pasientene ble } \mathbf{friske}}_{\text{effekt } \geq 0.85}$.

Basert på $\hat{p}$, kan vi nå *inferere* (= konkludere) at suksess-sannsynligheten er 0.88 for *alle* med den aktuelle sykdommen - ikke bare for de som faktisk ble testet.

Spørsmål:

1. Er $\hat{p}$ en *korrekt* estimator for p?
2. Er $\hat{p}$ en *god* estimator for p?
3. Hva er *øvre og nedre grenser* for intervallet p tilhører med tilnærmet sannsynlighet 95 %?

For å forstå dybden i disse spørsmålene, må vi først innse at $\hat{p}$ er en *stokastisk variabel*.
Tallet 0.88 er en realisering av estimatoren basert på akkurat de 100 tilfeldig utvalgte pasientene vi valgte. Dersom vi gjentok forsøksrekken med at *nytt* utvalg på 100 pasienter, hadde vi ganske sikkert fått et nytt estimat.

2) Andre tilnærming - $\hat{p}$ er en stokastisk variabel:

Estimatoren $\hat{p}$ er forbundet med *usikkerhet* og er således en stokastisk variabel, dvs. $\hat{p}$ er en funksjon av de stokastiske forsøksvariablene $\underbrace{X_1, X_2 \ldots X_n}_{0 \text{ eller } 1}$: ($n = 100$)

$$\overbrace{\hat{p}}^{\text{stok. var.}} = \hat{p}(\overbrace{X_1, X_2 \ldots X_n}^{\text{stok. var.}}) = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X} \qquad (6.5)$$

Fra observasjonene $x_1, x_2 \ldots x_n$ (tallverdier 0 eller 1) fra forsøksrekken får vi en *realisering* av estimatoren:

$$\hat{p}(\overbrace{x_1, x_2 \ldots x_n}^{0 \text{ eller } 1}) = \underbrace{\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i}_{\text{tall mellom 0 og 1}} = \frac{88}{100} = 0.88 \qquad (6.6)$$

som er samme svar som i den første tilnærmelsen i lign.(6.4).

Legg merke til at:

- x_i = $\underbrace{\text{observasjonen}}_{\text{tall}}$ tilhørende forsøk nr. i
- $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ er et tall mellom 0 og 1 siden $\overline{x}$ bare er gj.snittet mellom 0'ere og 1'ere

- X_i = den stokastiske forsøksvariabelen (som beskriver *sannsynlighetsmodellen* for forsøk nr. i)
- $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ gjennomsnitt av de stokastiske variablene

Svar på spørsmål:

1. Er $\hat{p}$ en *korrekt* estimator for p?

   Svar:

   En måte å si at en estimator er *korrekt* er dersom den forventede verdien til estimatoren er lik den sanne verdien:

$$\underline{\underline{E[\hat{p}]}} \;=\; E[\,\overline{X}\,] \;=\; E\Big[\,\frac{1}{n}\sum_{i=1}^{n} X_i\,\Big] \;=\; \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{n}{n}\,p \;=\; \underline{\underline{p}} \tag{6.7}$$

   siden, for en Bernoulli-fordeling, så er: [1]

$$E[X_i] \;=\; p \tag{6.11}$$

   En slik estimator $\hat{p}$ kalles en *forventningsrett* estimator.

■

[1] Fra lign.(5.15) på side 270 vet vi at både forsøksvariablene X_i og populasjonsvariablene X er Bernoulli-fordelt:

$$X_1,\, X_2,\, X_3,\, \ldots\,,\, X_n \;\overset{\text{i.i.d.}}{\sim}\; X \;\sim\; \text{Ber}[\,p\,] \tag{6.8}$$

For populasjonsvariablene er forventningen:

$$\underline{E[X]} \;=\; \sum_{i=0}^{1} x_i P(X = x_i) \;=\; 0\cdot(1-p) + 1\cdot p \;=\; \underline{p} \tag{6.9}$$

og tilsvarende for forsøksvariablene:

$$E[X_i] \;=\; p \tag{6.10}$$

2. Er $\hat{p}$ en *god* estimator for p?

   Svar:

   En estimator $\hat{p}$ er *god* dersom [2]

$$\boxed{\left| P_{\hat{p}}(A) - P_p(A) \right|} \qquad (6.12)$$

   er så liten som mulig for alle begivenheter $A \subset V$. Lign. (6.12) sier at sannsynlighetsloven $P_{\hat{p}}$ vi får fra estimatet $\hat{p}$ er så "nær" den virkelige sannsynlighetsloven P_p som mulig.

   Dette målet gir grunnlaget for å kunne si om en estimator er *bedre* enn en annen. Vi kommer ikke inn på dette området her.

■

[2] Streken | betyr absoluttverditegn, f.eks. $|0.4 - 0.6| = 0.2$.

3. Hva er *øvre og nedre grenser* for intervallet som p tilhører med tilnærmet sannsynlighet 95 %?

Svar: [3]

Intervallet ($n = 100$)

$$\boxed{\left[\ LB_n^p\ ,\ UB_n^p\ \right]} \qquad (6.13)$$

inneholder den sanne sannsynligheten p med tilnærmet sannsynlighet minst 95 %, hvor: [4]

$$LB_n^p = 0.8163 \qquad (6.14)$$

$$UB_n^p = 0.943 \qquad (6.15)$$

Intervallet i lign.(6.13) kalles et *asymptotisk* 95 %-*konfidensintervall*. [5]

■

I neste avsnitt formaliseres disse begrepene.

[3] LB = "lower bound" og UB = "upper bound".

[4] At sannsynligheten er *tilnærmet* 95 % kommer av sentralgrenseteoremet, som sier at $\hat{p} = \overline{X}$ er *tilnærmet* normalfordelt.

[5] Vi skal senere vise hvordan man finner dette intervallet.

6.2 Estimatorer

Felles antagelser: (for definisjoner og setninger i avsnitt 6.2)

Anta at:

$$X_1,\, X_2,\, X_3,\, \ldots\, , X_n \overset{\text{i.i.d.}}{\sim} X \sim P_\theta(X = x) \tag{6.16}$$

for $\theta \in \Theta$.

Definisjon: (estimator)

En estimator $\hat{\theta}$ for den sanne parameteren $\theta \in \Theta$ er en funksjon av forsøksvariablene:

$$\boxed{\hat{\theta} \;=\; \hat{\theta}(X_1, X_2 \ldots X_n)} \tag{6.17}$$

■

Definisjon: (forventningsrett estimator)

En estimator $\hat{\theta}$ sies å være forventnings**rett** dersom

$$\boxed{E[\,\hat{\theta}\,] \;=\; \theta} \tag{6.18}$$

hvor θ den sanne parameteren $\theta \in \Theta$.

I motsatt fall er den forventningskjev.

■

Kommentarer:

- At en estimator er forventningsrett betyr at dersom forsøket gjentas mange ganger vil estimatoren i gjennomsnitt, i det lange løp, gi **rett** verdi.
- At en estimator er forventningskjev betyr at dersom forsøket gjentas mange ganger vil estimatoren i gjennomsnitt, i det lange løp, gi **gal** verdi, (dvs. en systematisk feil ved å bruke en estimator som ikke er forventningsrett).

Eksempel: (estimator for suksess-sannsynligheten p - legemiddel)

Vi viste i lign.(6.7) at estimatoren:

$$\boxed{\hat{p}(X_1, \ldots, X_{100}) \;=\; \overline{X}} \qquad (6.19)$$

er en forventningsrett estimator for p siden

$$E[\,\hat{p}(X_1, \ldots, X_{100})\,] \overset{\text{lign.(6.7)}}{=} p \qquad (6.20)$$

■

Eksempel: (estimator - Y_i, dvs. effekt)

Anta at:

$$Y_1,\, Y_2,\, Y_3,\, \ldots\, , Y_n \overset{\text{i.i.d.}}{\sim} Y \sim N[\,\mu\,,\,\sigma\,] \tag{6.21}$$

hvor de stokastiske forsøksvariablene $\overbrace{Y_i}^{i=1,2,3\ldots,100}$ for pasient nr. i er: (se lign.(5.16) side 271)

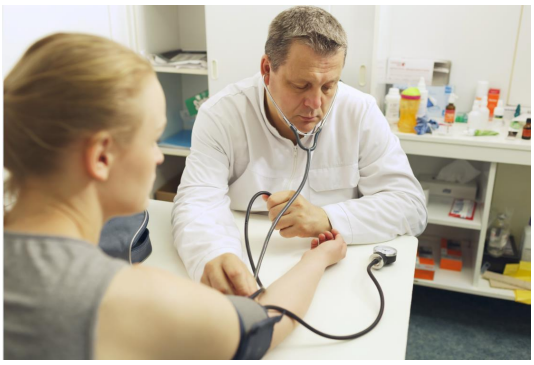
$$Y_i = \underbrace{\textbf{effekten} \text{ av legemiddelet}}_{\text{tall mellom 0 og 1, se tabell 5.2}} \text{ for forsøkspasient nr. } i \tag{6.22}$$

og de stokastisk populasjonsvariabelene er:

$$Y = \textbf{effekten} \text{ av legemiddelet for en } \underbrace{\text{tilfeldig valgt pasient i populasjonen}}_{\text{alle som har sykdommen, ikke bare utvalget } n=100} \tag{6.23}$$

Lign.(6.21) betyr blant annet at $E[Y] = E[Y_i] = \mu$ og at $\sigma^2[Y] = \sigma^2[Y_i] = \sigma^2$.

Foto: Colourbox



Figur 6.5: Forsøk.

Vi ønsker å finne *forventningsrette* estimatorer $\hat{\mu}$ og $\hat{\sigma}^2$.
Det er naturlig å foreslå følgende kandidater:

$$\hat{\mu} \;=\; \overline{Y} \;=\; \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad (6.24)$$

$$\hat{\sigma}^2 \;=\; S_{y,n}^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \qquad (6.25)$$

hvor $n = 100$. [6]

a) Vis at $\hat{\mu}$ er forventningsrett.

b) Vis at $\hat{\sigma}^2$ ikke er forventningsrett.

c) Bestem på bakgrunn av oppgave **b** en estimator som er forventningsrett.
Kjenner du igjen denne fra kapittel 5?

[6] Legg merke til at det står $\frac{1}{n}$ i lign.(6.25). Det er fordi det er naturlig å foreslå et gjennomsnitt. Lign.(5.23) på side 276, derimot, har prefaktoren $\frac{1}{n-1}$ for den empiriske variansen s_x^2.

Løsning:

a) Forventningen av estimatoren $\hat{\mu}$ i lign.(6.24): [7]

$$\underline{E[\,\hat{\mu}\,]} \;=\; E[\,\overline{Y}\,] \tag{6.26}$$

$$=\; E\Big[\,\frac{1}{n}\sum_{i=1}^{n} Y_i\,\Big] \tag{6.27}$$

$$=\; \frac{1}{n}\sum_{i=1}^{n} E\big[\,Y_i\,\big] \;=\; \frac{\not{n}\,\overbrace{E[Y_i]}^{=\,\mu}}{\not{n}} \;=\; \underline{\mu} \tag{6.28}$$

siden vi har antatt at $Y_i \sim N[\mu, \sigma]$, dvs.

$$E[Y_i] \;=\; \mu \tag{6.29}$$

Lign.(6.28) viser at forventningen av estimatoren $\hat{\mu}$ er den sanne forventningen μ.

Estimatoren $\hat{\mu}$ er altså forventningsrett.

[7] Husk at $\hat{\mu}$ er en stokastisk variabel. Derfor gir det mening å ta forventningsverdien av den.

b) Forventningen av estimatoren $\hat{\sigma}^2$ i lign.(6.25): [8]

$$\underline{E[\,\hat{\sigma}^2\,]} = E[\,S^2_{y,n}\,] \tag{6.30}$$

$$= E\Big[\,\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2\,\Big] \tag{6.31}$$

$$= E\Big[\,\frac{1}{n}\sum_{i=1}^{n}\Big(\,Y_i^2 - 2Y_i\overline{Y} + \overline{Y}^2\,\Big)\,\Big] \tag{6.32}$$

.
.
. and then a miracle occurs
.
.

$$= \underline{\frac{n-1}{n}\,\sigma^2} \tag{6.33}$$

som viser at forventningen av estimatoren $\hat{\sigma}^2$ ikke er den sanne variansen σ^2.

Estimatoren $S^2_{y,n}$ er altså ikke forventningsrett, dvs. den er forventningskjev.

[8] Husk at $\hat{\sigma}^2$ er en stokastisk variabel. Derfor gir det mening å ta forventningsverdien av den.

c) I lys av lign.(5.23) på side 276 så ser vi på følgende estimator:

$$\boxed{\hat{\sigma}^2 \;=\; S_y^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \qquad (6.34)$$

som har prefaktoren $\frac{1}{n-1}$. Fra lign.(6.25) ser vi at

$$\underline{S_y^2} \;=\; \frac{n}{n-1}\,\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \;\overset{\text{lign.(6.25)}}{=}\; \underline{\frac{n}{n-1}\,S_{y,n}^2} \qquad (6.35)$$

Forventningen av den stokastiske variabelen $\hat{\sigma}^2$ i lign.(6.25):

$$\begin{aligned}
\underline{E[\,\hat{\sigma}^2\,]} \;&=\; E[\,S_y^2\,] && (6.36)\\
&=\; E\Big[\,\frac{n}{n-1}\,S_{y,n}^2\,\Big] && (6.37)\\
&=\; \frac{n}{n-1}\,\overbrace{E\Big[\,S_{y,n}^2\,\Big]}^{=\,\frac{n-1}{n}\,\sigma^2} && (6.38)\\
&=\; \frac{\cancel{n}}{\cancel{n-1}}\,\frac{\cancel{n-1}}{\cancel{n}}\,\sigma^2 && (6.39)\\
&=\; \underline{\sigma^2} && (6.40)
\end{aligned}$$

som viser at forventningen av estimatoren $\hat{\sigma}^2$ i lign.(6.34) er den sanne forventningen σ^2. Estimatoren $\hat{\sigma}^2$ er altså forventningsrett.

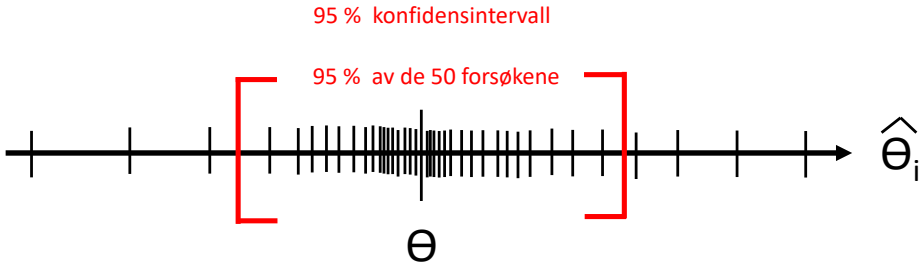
Ja, vi ser at realiseringen av estimatoren $\hat{\sigma}^2$ i lign.(6.34) er den empiriske variansen s_x^2 som definert i lign.(5.23) på side 276.

■

6.3 Konfidensintervaller

I forrige avsnitt definerte vi begrepet $\overbrace{\textit{estimator}\ \hat{\theta}}^{\text{stok. var.}}$ for $\overbrace{\theta}^{\textit{tall}}$.

En slik estimator er et *punktestimat*, dvs. det sier ikke noe om hvor mye $\hat{\theta}$ *varierer* rundt θ.
Hvis vi gjentar forsøksrekken 50 ganger, får vi 50 forskjellige punktestimat for θ.
Hvor stor er spredningen til disse 50 realiseringene $\hat{\theta}$?



Figur 6.6: Spredning rundt θ.

Konfidensintervallet gir oss svaret på spørsmålet ovenfor.
Vi antar, som i for forrige avsnitt om estimatorer, at vi har gjennomført en statistisk forsøksrekke.

Felles antagelser: (for definisjoner og setninger i avsnitt 6.3)

Anta at:

$$X_1,\, X_2,\, X_3,\, \ldots\, , X_n \overset{\text{i.i.d.}}{\sim} X \sim P_\theta(X = x) \tag{6.41}$$

for $\theta \in \Theta$ med verdimengde V.

Definisjon: ($(1-\alpha)\,100\,\%$-konfidensintervall for θ)

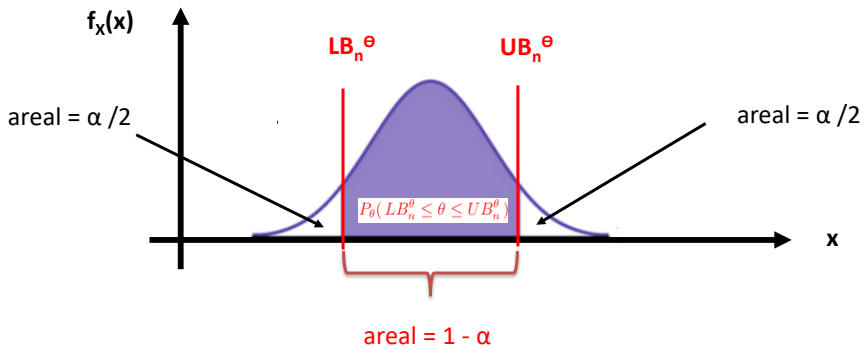
Et $(1-\alpha)\,100\,\%$-konfidensintervall for θ er et *stokastisk intervall*

$$\boxed{[\,LB_n^\theta\,,\, UB_n^\theta\,]} \tag{6.42}$$

hvor LB_n^θ og UB_n^θ er fastsatt slik at sannsynligheten for at θ er inneholdt i dette intervallet er $(1-\alpha)\ 100\,\%$, dvs.:

$$\boxed{P(\,LB_n^\theta \le \theta \le UB_n^\theta\,) \;=\; 1-\alpha} \tag{6.43}$$

■



Figur 6.7: Konfidensintervall.

Kommentarer:

- Tallet α kalles

$$\boxed{\alpha \;=\; \text{signifikansnivået}}$$

og er en størrelse som er valgt av de som gjennomfører analysen. Som regel velges en *lav* verdi for α, dvs. typisk er $\alpha = 0.05$ eller lavere. Med $\alpha = 0.05$ fås et 95 % konfidensintervall, dvs. at intervallet inneholder θ med 95 % sannsynlighet.

- LB_n^θ står for "Lower Bound" eller nedre grense på norsk, og er en funksjon av forsøksvariablene:

$$\underbrace{LB_n^\theta \;=\; LB_n^\theta(X_1, \ldots X_n)}_{\text{stokastisk variabel}}$$

LB_n^θ er derfor en stokastisk variabel. Legg også merke til at LB_n^θ er en funksjon av størrelsen på utvalget, dsv. n.

- UB_n^θ står for "Upper Bound" eller øvre grense på norsk, og er en funksjon av forsøksvariablene:

$$\underbrace{UB_n^\theta \;=\; UB_n^\theta(X_1, \ldots X_n)}_{\text{stokastisk variabel}}$$

UB_n^θ er derfor en stokastisk variabel. Legg også merke til at UB_n^θ er en funksjon av størrelsen på utvalget, dvs. n.

Sannsynligheten i lign.(6.43) er ofte vanskelig å beregne nøyaktig.
Typisk vil både LB_n^θ og UB_n^θ være funksjoner av gjennomsnittet $\overline{X} = \frac{1}{n}\Big(X_1 + X_2 + X_3 + ... + X_n \Big)$.
Siden forsøksvariablene X_i er i.i.d. kan vi benytte *sentralgrenseteoremet* fra kapittel 4:
(typisk $n \gtrsim 30$, se lign.(4.172) på side 248)

$$\lim_{n\to\infty} \frac{\overline{X} - E[\overline{X}]}{\sigma[\overline{X}]} \;\sim\; N[0,1] \tag{6.44}$$

Ved hjelp av lign.(6.44) kan tilnærmede uttrykk for LB_n^θ og UB_n^θ bestemmes slik at sannsynligheten i lign.(6.43) blir mer korrekt desto større n blir (utvalget).

Vi får det som kalles et *asymptotisk* $(1-\alpha)\,100\,\%$-konfidensintervall for θ.

Definisjon: (asymptotisk $(1-\alpha)\,100\,\%$-konfidensintervall for θ)

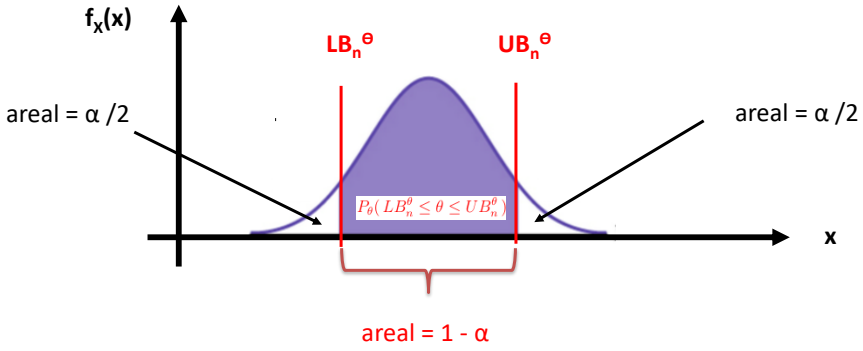
Et asymptotisk $(1-\alpha)\,100\,\%$-konfidensintervall for θ er et *stokastisk intervall*

$$\boxed{[\,LB_n^\theta\,,\,UB_n^\theta\,]} \qquad (6.45)$$

hvor LB_n^θ og UB_n^θ er fastsatt slik at sannsynligheten for at θ er inneholdt i dette intervallet er $(1-\alpha)\,100\,\%$ når $n \to \infty$, dvs.:

$$\boxed{\lim_{n\to\infty} P(\,LB_n^\theta \le \theta \le UB_n^\theta\,) \;=\; 1-\alpha} \qquad (6.46)$$

■



Figur 6.8: Konfidensintervall.

Eksempel: (asymptotisk 95 %-konfidensintervall - sannsynlighet p)

Anta at:

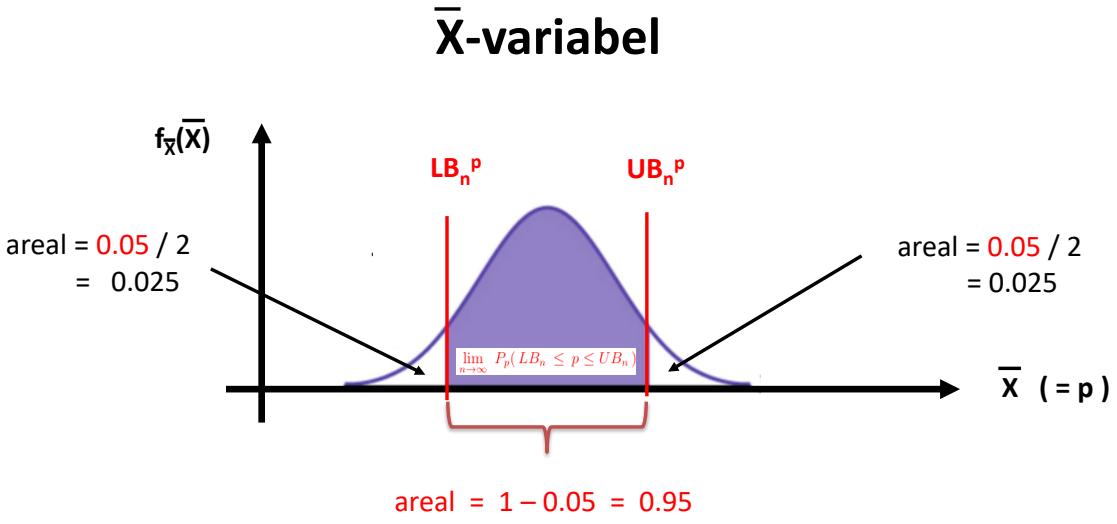
$$X_1,\, X_2,\, X_3,\, \ldots\, , X_n \overset{\text{i.i.d.}}{\sim} X \sim \text{Ber}[p] \tag{6.47}$$

hvor de stokastiske forsøksvariablene X_i beskriver resultatene fra forsøkene med legemiddelet, dvs.:

$$X_i = \begin{cases} 1 & , \overbrace{\text{dersom pasient nr. } i \text{ blir frisk}}^{=\, p} \\ 0 & , \underbrace{\text{hvis ikke}}_{=\, 1-p} \end{cases} \tag{6.48}$$

hvor

$$p = \text{suksess-sannsynlighet (ukjent } p \text{)} \tag{6.49}$$



Figur 6.9: 95 %-konfidensintervall.

Finn et asymptotisk konfidensintervall

$$[\, LB_n^p \,,\, UB_n^p \,] \tag{6.50}$$

for den forventningsrette estimatoren $\hat{p}$, dvs.:

$$\hat{p}(X_1, \ldots, X_{100}) \overset{\text{lign.(6.19)}}{=} \overline{X} \tag{6.51}$$

med signifikansnivå $\alpha = 0.05$.

Løsning:

Definisjonen av et asymptotisk konfidensintervall er gitt ved lign.(6.46) med $\theta = p$:

$$\boxed{\lim_{n\to\infty} P(\, LB_n^p \leq p \leq UB_n^p \,) = 1-\alpha} \qquad (6.52)$$

Fra lign.(6.19) vet vi da at: [9]

$$\hat{p} = \overline{X} \qquad (6.54)$$

er en forventningsrett estimator for den ukjente p.
Alle X_i i lign.(6.54) oppfyller i.i.d. (med $X_i \sim \text{Ber}[p]$). Fra sentralgrenseteoremet vet vi da at:

$$Z = \frac{\overline{X} - E[\overline{X}]}{\sigma[\overline{X}]} \overset{n\to\infty}{\sim} N\big[\,0\,,\,1\,\big] \qquad (6.55)$$

Med $n = 100$ så er n stor nok til at Z i lign.(6.55) med god tilnærmelse er en normalfordeling.
Vi må finne $E[\overline{X}]$ og $\sigma[\overline{X}]$ for å standardisere.

[9] Husk at:

$$\overline{X} = \frac{X_1 + X_2 + ...X_n}{n} \qquad (6.53)$$

Fra lign.(6.7) vet vi at

$$E[\,\hat{p}\,] \quad = \quad E[\,\overline{X}\,] \overset{\text{lign.(6.7)}}{=} p \tag{6.56}$$

altså estimatoren $\hat{p}$ er forvetningsrett. På tilsvarende måte kan man vise at: [10]

$$Var[\,\hat{p}\,] \quad = \quad Var[\,\overline{X}\,] \; = \; \frac{p(1-p)}{n} \tag{6.64}$$

Generelt er $\sigma[X] = \sqrt{Var[X]}$, dermed:

$$\sigma[\,\hat{p}\,] \quad = \quad \sigma[\overline{X}] \; = \; \sqrt{\frac{p(1-p)}{n}} \tag{6.65}$$

[10] Fra lign.(5.15) på side 270 vet vi at både forsøksvariablene X_i og populasjonsvariablene X er Bernoulli-fordelt:

$$X_1,\, X_2,\, X_3,\, \ldots\,, X_n \overset{\text{i.i.d.}}{\sim} X \;\sim\; \text{Ber}[p] \tag{6.57}$$

Variansen av $\hat{p} \;=\; \overline{X}$:

$$\underline{Var[\,\hat{p}\,]} \quad = \quad Var[\,\overline{X}\,] \tag{6.58}$$

$$= \quad Var\Big[\; \frac{X_1 + X_2 + X_3 + ... + X_n}{n} \;\Big] \tag{6.59}$$

$$\overset{\text{i.i.d.}}{=} \quad \frac{1}{n^2}\Big(\; Var[X_1] + Var[X_2] + Var[X_3] + ... + Var[X_n] \;\Big) \tag{6.60}$$

$$= \quad \frac{\cancel{n} Var[X_i]}{n^{\cancel{2}}} \quad = \quad \underline{\frac{p(1-p)}{n}} \tag{6.61}$$

hvor vi har brukt at $Var[X_i] = Var[X]$ med

$$\underline{Var[X]} \quad = \quad \sum_{i=0}^{1} \big(\; x_i - E[X] \;\big) P(X = (x_i)) \tag{6.62}$$

$$= \quad (\,0-p\,)^2(1-p) \;+\; (\,1-p\,)^2 p \;=\; \underline{p(1-p)} \tag{6.63}$$

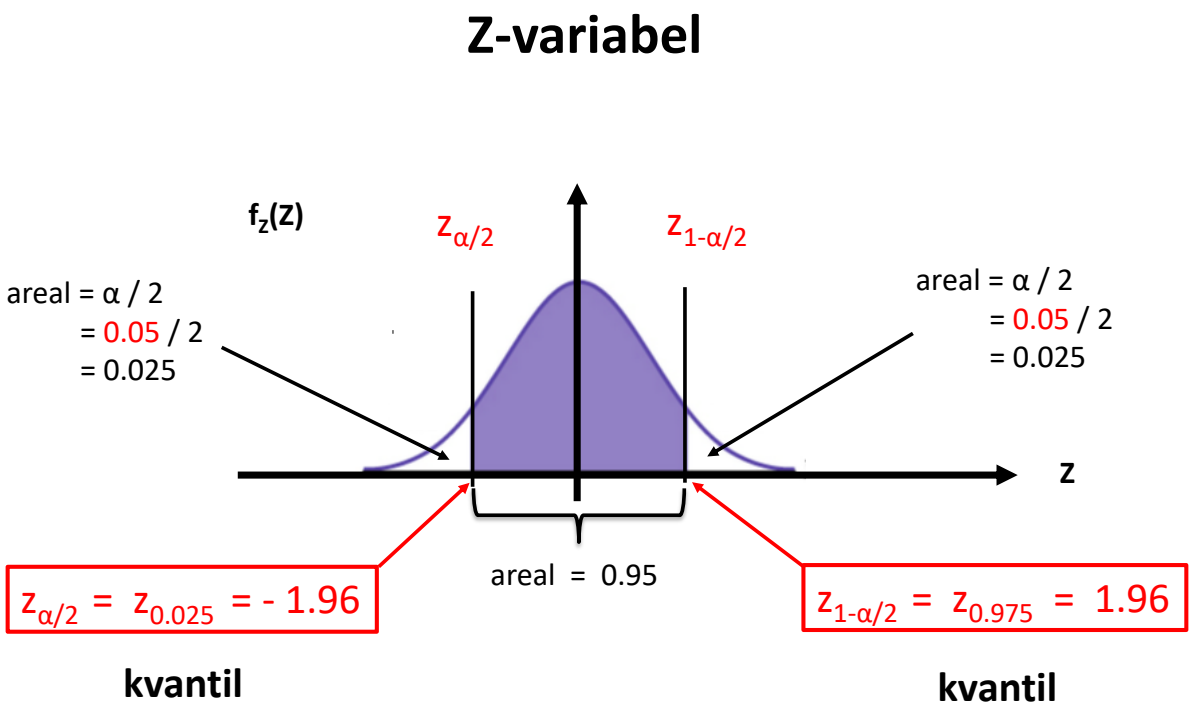
Variabelen Z i lign.(6.55) er da:

$$\underline{Z} \;=\; \frac{\overline{X} - E[\overline{X}]}{\sigma[\overline{X}]} \;=\; \frac{\overline{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \;\overset{n\to\infty}{\sim}\; N[0,1] \tag{6.66}$$

Vi finner kvantilene i figur 6.10: [11]

$$z_{\alpha/2} \;=\; z_{0.025} \;=\; -1.96 \quad , \quad z_{1-\alpha/2} \;=\; z_{0.975} \;=\; 1.96 \tag{6.67}$$

via omvendt tabelloppslag fra side 208.



Figur 6.10: Kvantil.

[11] Absoluttverdiene til $z_{\alpha/2}$ og $z_{1-\alpha/2}$ er like fordi N-fordelingen er symmetrisk.

Fra definisjonen i lign.(6.52):

$$\lim_{n\to\infty} P(\, LB_n^p \,\leq\, p \leq UB_n^p \,) \;\;=\;\; 1-\alpha \tag{6.68}$$

Siden estimatoren $\hat{p}$ er normalfordelt for $n \to \infty$, se lign.(6.54) og (6.55), så standardiserer vi:

$$\lim_{n\to\infty} P\big(\, z_{\alpha/2} \,\leq\, Z \leq z_{1-\alpha/2} \,\big) \;\;=\;\; 1-\alpha \tag{6.69}$$

hvor Z er gitt ved lign.(6.66):

$$\lim_{n\to\infty} P\left(\; z_{\alpha/2} \;\leq\; \frac{\overline{X}-p}{\sqrt{\frac{p(1-p)}{n}}} \;\leq\; z_{1-\alpha/2} \; \right) \;=\; 1-\alpha \tag{6.70}$$

$$\Updownarrow \qquad \text{(algebra)}$$

$$\lim_{n\to\infty} P\left(\; \underbrace{\overline{X} - \frac{z_{1-\alpha/2}}{\sqrt{n}}\sqrt{p(1-p)}}_{=\, LB_n^p} \;\leq\; p \;\leq\; \underbrace{\overline{X} - \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{p(1-p)}}_{=\, UB_n^p} \; \right) \;=\; 1-\alpha \tag{6.71}$$

og hvor vi definerer nedre og øvre grenser:

$$LB_n^p \;\;=\;\; \overline{X} \;-\; \frac{z_{1-\alpha/2}}{\sqrt{n}}\sqrt{p(1-p)} \tag{6.72}$$

$$UB_n^p \;\;=\;\; \overline{X} \;-\; \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{p(1-p)} \tag{6.73}$$

Problem:

Vi kan ikke bestemme en realisering av LB_n^p og UB_n^p siden de er avhengige av den ukjente suksess-sannsynligheten p.

Løsning:

Vi kan bytte ut p i lign.(6.72) og (6.73) med estimatoren $\hat{p} = \overline{X}$.
Det viser seg at denne tilnærmingen også oppfyller sentralgrenseteoremet. [12]

Med $\hat{p} = \overline{X}$ innsatt for p i lign.(6.72) og (6.73):

$$LB_n^p = \overline{X} - \frac{z_{1-\alpha/2}}{\sqrt{n}}\sqrt{\overline{X}(1-\overline{X})} \tag{6.74}$$

$$UB_n^p = \overline{X} - \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{\overline{X}(1-\overline{X})} \tag{6.75}$$

hvor

$$z_{\alpha/2} = \text{nedre kvantil til } N[0,1]\text{-fordelingen} \tag{6.76}$$

$$z_{1-\alpha/2} = \text{øvre kvantil til } N[0,1]\text{-fordelingen} \tag{6.77}$$

[12] Dette kommer som en følge av den såkalte store talls lov, som vi ikke skal komme inn på her.

Intervallet

$$\boxed{[\, LB_n^p \,,\, UB_n^p \,] \;=\; \left[\; \overline{X} \;-\; \frac{z_{1-\alpha/2}}{\sqrt{n}} \sqrt{\overline{X}(1-\overline{X})} \;,\; \overline{X} \;-\; \frac{z_{\alpha/2}}{\sqrt{n}} \sqrt{\overline{X}(1-\overline{X})} \;\right]} \tag{6.78}$$

er dermed et *asymptotisk* $(1-\alpha)\,100\,\%$-konfidensintervall for suksess-sannsynligheten p.

Med $\alpha = 0.05$ så finner vi ved tabelloppslag:

$$z_{\alpha/2} \;=\; z_{0.025} \;=\; -1.96 \tag{6.79}$$

$$z_{1-\alpha/2} \;=\; z_{0.975} \;=\; 1.96 \tag{6.80}$$

Fra dataene $x_1, x_2 \ldots x_{100}$ fra forsøksrekken får vi en *realisering* (små $x_1, x_2, \ldots, x_{100}$) av den nedre og øvre grensen: ($n = 100$)

$$\underline{LB_n^p(x_1, x_2, \ldots, x_n)} = \overline{x} - \frac{z_{1-\alpha/2}}{\sqrt{n}} \sqrt{\overline{x}(1-\overline{x})} \tag{6.81}$$

$$= 0.88 - \frac{1.96}{\sqrt{100}} \sqrt{0.88(1-0.88)} \tag{6.82}$$

$$= \underline{0.8163} \tag{6.83}$$

$$\underline{UB_n^p(x_1, x_2, \ldots, x_n)} = \overline{x} - \frac{z_{\alpha/2}}{\sqrt{n}} \sqrt{\overline{x}(1-\overline{x})} \tag{6.84}$$

$$= 0.88 - \frac{(-1.96)}{\sqrt{100}} \sqrt{0.88(1-0.88)} \tag{6.85}$$

$$= \underline{0.9437} \tag{6.86}$$

som gir realiseringen

$$\underline{\underline{[\ LB_n^p\ ,\ UB_n^p\ ]}} = [\ 0.8163\ ,\ 0.9437\ ] \tag{6.87}$$

av det asymptotiske 95 %-konfidensintervallet for sukess-sannsynligheten p.

■

Eksempel: (asymptotisk 95 %-konfidensintervall - effekten $\hat{\mu} = \overline{Y}$)

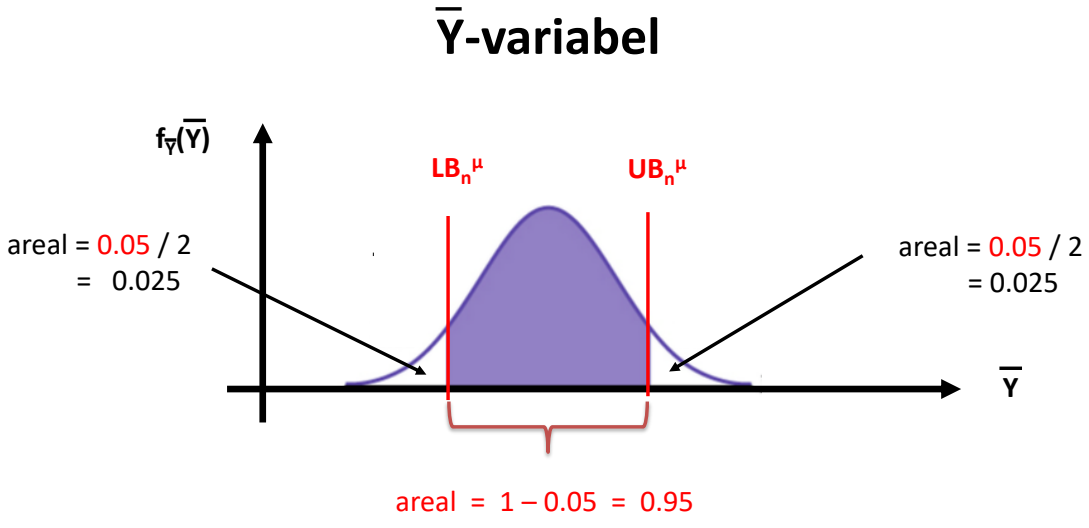
Anta at:

$$Y_1, Y_2, Y_3, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} Y \sim N[\mu, \sigma] \tag{6.88}$$

hvor de stokastiske forsøksvariablene Y_i beskriver effekten for pasient nr. i:

$$Y_i \overset{\text{lign.(5.16)}}{=} \text{effekten for forsøkspasient nr. } i \tag{6.89}$$

Lign.(6.88) betyr blant annet at $E[Y] = E[Y_i] = \mu$ og at $\sigma^2[Y] = \sigma^2[Y_i] = \sigma^2$.



Figur 6.11: 95 %-konfidensintervall.

Finn et asymptotisk konfidensintervall

$$[\, LB_n^{\mu}\,,\, UB_n^{\mu}\,] \tag{6.90}$$

for den sanne forventningsverdien μ med signifikansnivå $\alpha = 0.05$.

Løsning:

Definisjonen av et asymptotisk konfidensintervall er gitt ved lign.(6.46) med $\theta = \mu$:

$$\boxed{\lim_{n\to\infty} P_\mu(LB_n^\mu \leq \mu \leq UB_n^\mu) = 1-\alpha} \qquad (6.91)$$

Alle Y_i i lign.(6.88) oppfyller i.i.d. (med $Y_i \sim \mathrm{N}[\mu, \sigma]$).
Sum normalfordelinger er forsatt normalfordelt: (se lign.(4.180) side 253)

$$\overline{Y} \sim N\big[\, E[\overline{Y}]\,,\, \sigma[\overline{Y}]\,\big] \qquad (6.92)$$

Vi må standardisere. [13]

[13]Estimatoren $\hat{\mu}$ er forventningsrett, dvs.:

$$E[\,\hat{\mu}\,] = E[\,\overline{Y}\,] = \mu \qquad (6.93)$$

altså estimatoren $\hat{\mu}$ er forvetningsrett. Variansen til $\hat{\mu}$ er:

$$Var[\,\hat{\mu}\,] = Var[\,\overline{Y}\,] = \frac{\sigma^2}{n} \qquad (6.94)$$

Generelt er $\sigma[Y] = \sqrt{Var[Y]}$, dermed:

$$\sigma[\,\hat{\mu}\,] = \sigma[\overline{Y}] = \frac{\sigma}{\sqrt{n}} \qquad (6.95)$$

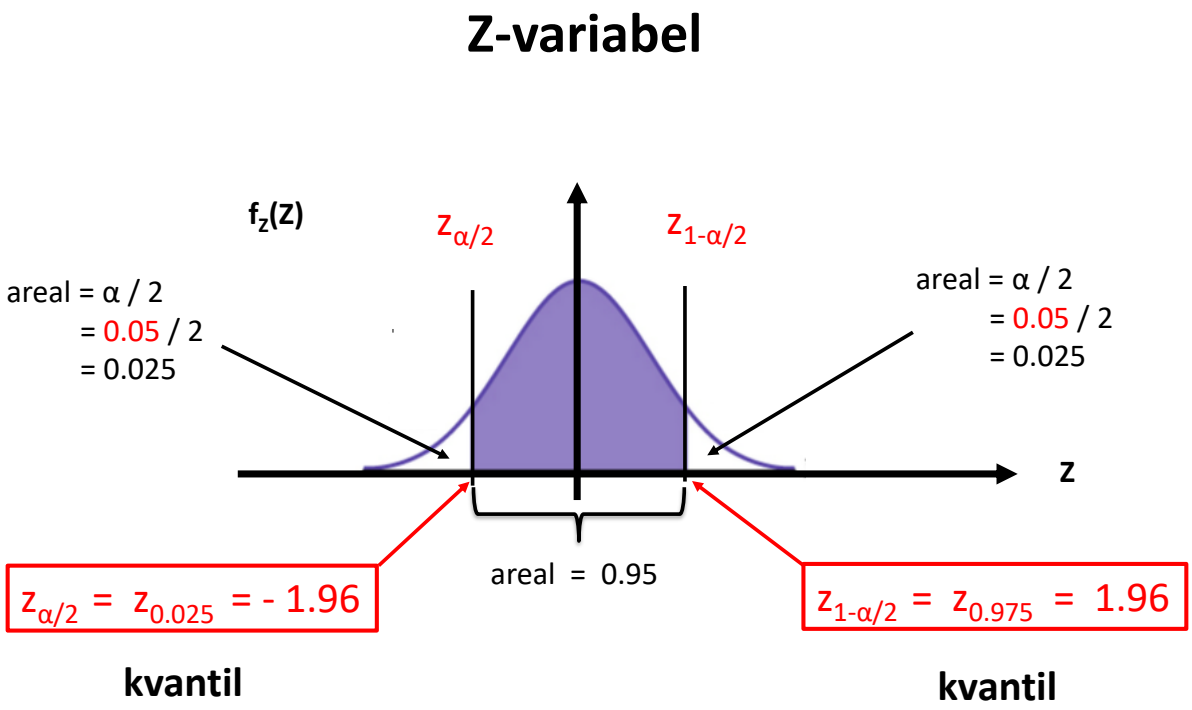
Standardiserer:

$$\underline{Z} \;=\; \frac{\overline{X} - E[\overline{Y}]}{\sigma[\overline{Y}]} \;=\; \underline{\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}} \;\overset{n\to\infty}{\sim}\; N[0,1] \tag{6.96}$$

Vi finner kvantilene figur 6.12: [14]

$$z_{\alpha/2} \;=\; z_{0.025} \;=\; -1.96 \quad , \quad z_{1-\alpha/2} \;=\; z_{0.975} \;=\; 1.96 \tag{6.97}$$

via omvendt tabelloppslag fra side 208.



Figur 6.12: Kvantil.

[14] Absoluttverdiene til $z_{\alpha/2}$ og $z_{1-\alpha/2}$ er like fordi N-fordelingen er symmetrisk.

Fra definisjonen i lign.(6.91):

$$\lim_{n\to\infty} \; P\big(\, LB_n^{\mu} \;\leq\; \mu \leq UB_n^{\mu}\,\big) \;\;=\;\; 1-\alpha \tag{6.98}$$

Siden estimatoren $\hat{\mu}$ er normalfordelt, se lign.(6.92), så standardiserer vi:

$$\lim_{n\to\infty} \; P\big(\, z_{\alpha/2} \;\leq\; Z \leq z_{1-\alpha/2}\,\big) \;\;=\;\; 1-\alpha \tag{6.99}$$

hvor Z er gitt ved lign.(6.96):

$$\lim_{n\to\infty} \; P\left(\; z_{\alpha/2} \;\leq\; \frac{\overline{Y}-\mu}{\frac{\sigma}{\sqrt{n}}} \;\leq\; z_{1-\alpha/2} \; \right) \;=\; 1-\alpha \tag{6.100}$$

$$\Updownarrow \qquad \text{(algebra)}$$

$$\lim_{n\to\infty} \; P\left(\; \underbrace{\overline{Y} - \frac{z_{1-\alpha/2}}{\sqrt{n}}\,\sigma}_{=\; LB_n^{\mu}} \;\leq\; \mu \;\leq\; \underbrace{\overline{Y} - \frac{z_{\alpha/2}}{\sqrt{n}}\,\sigma}_{=\; UB_n^{\mu}} \; \right) \;=\; 1-\alpha \tag{6.101}$$

og hvor vi definerer nedre og øvre grenser:

$$LB_n^{\mu} \;\;=\;\; \overline{Y} \;-\; \frac{z_{1-\alpha/2}}{\sqrt{n}}\,\sigma \tag{6.102}$$

$$UB_n^{\mu} \;\;=\;\; \overline{Y} \;-\; \frac{z_{\alpha/2}}{\sqrt{n}}\,\sigma \tag{6.103}$$

Problem:

Vi kan ikke bestemme en realisering av LU_n^μ og UB_n^μ ovenfor siden vi er avhengige av det ukjente standardavviket σ.

Løsning:

Vi kan bytte ut σ i lign.(6.102) og (6.103) med estimatoren $\hat{\sigma} = S_y$, jfr. lign.(6.34) på side 311:

$$\hat{\sigma}^2 = S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \tag{6.104}$$

Det viser seg at denne tilnærmingen også oppfyller sentralgrenseteoremet. [15]

Med $\hat{\sigma} = S_y$ innsatt for σ i lign.(6.102) og (6.103):

$$LB_n^\mu = \overline{Y} - \frac{z_{1-\alpha/2}}{\sqrt{n}} S_y \tag{6.105}$$

$$UB_n^\mu = \overline{Y} - \frac{z_{\alpha/2}}{\sqrt{n}} S_y \tag{6.106}$$

[15] Dette kommer som en følge av den såkalte store talls lov, som vi ikke skal komme inn på her.

Intervallet

$$\boxed{[\, LB_n^{\mu}\, ,\, UB_n^{\mu}\,] \;=\; \left[\; \overline{Y} \;-\; \frac{z_{1-\alpha/2}}{\sqrt{n}}\, S_y \;,\; \overline{Y} \;-\; \frac{z_{\alpha/2}}{\sqrt{n}}\, S_y \;\right]} \tag{6.107}$$

er dermed et *asymptotisk* $(1-\alpha)\,100\,\%$-konfidensintervall for μ.

Med $\alpha = 0.05$ så finner vi ved tabelloppslag:

$$z_{\alpha/2} \;=\; z_{0.025} \;=\; -1.96 \tag{6.108}$$

$$z_{1-\alpha/2} \;=\; z_{0.975} \;=\; 1.96 \tag{6.109}$$

Fra dataene $y_1, y_2 \ldots y_{100}$ fra forsøksrekken (se tabell 5.2 side 271) samt $s_y = 0.0413$ fra tabell 5.5 side 278 får vi en *realisering* (små $y_1, y_2, \ldots, y_{100}$) av den nedre og øvre grensen: ($n = 100$)

$$\underline{LB_n^{\mu}(y_1, y_2, \ldots, y_n)} = \overline{y} - \frac{z_{1-\alpha/2}}{\sqrt{n}} s_y \tag{6.110}$$

$$= 0.8967 - \frac{1.96}{\sqrt{100}} \cdot 0.0413 \tag{6.111}$$

$$= \underline{0.8886} \tag{6.112}$$

$$\underline{UB_n^{\mu}(y_1, y_2, \ldots, y_n)} = \overline{y} - \frac{z_{\alpha/2}}{\sqrt{n}} s_y \tag{6.113}$$

$$= 0.8967 - \frac{(-1.96)}{\sqrt{100}} \cdot 0.0413 \tag{6.114}$$

$$= \underline{0.9048} \tag{6.115}$$

som gir realiseringen

$$\underline{\underline{[\ LB_n^{\mu}\ ,\ UB_n^{\mu}\ ]}} = [\ 0.8886\ ,\ 0.9048\ ] \tag{6.116}$$

av det asymptotiske 95 %-konfidensintervallet for μ.

■

Eksempel: (asymptotisk 95 %-konfidensintervall - standardavvik $\hat{\sigma} = S_y$ til effekten Y)

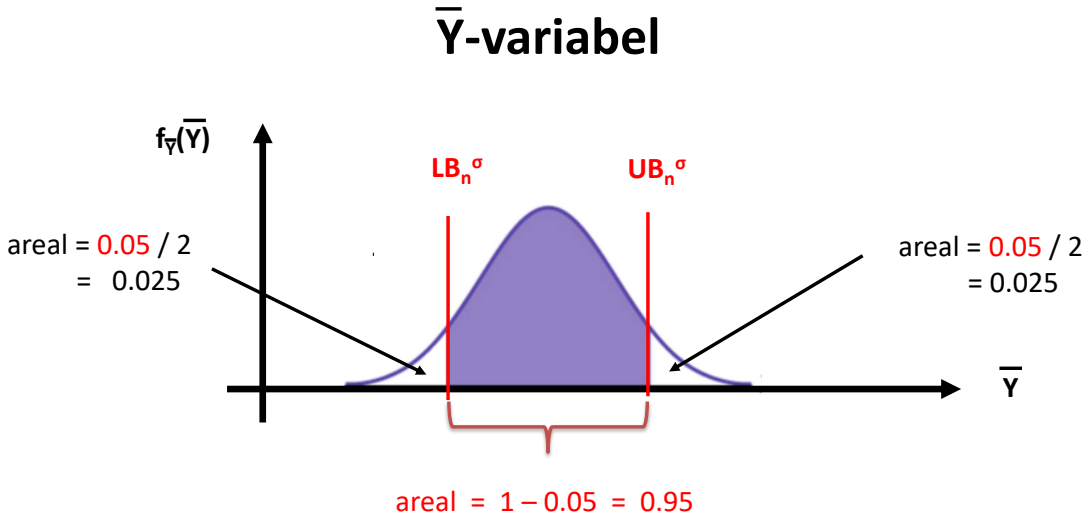
Anta at:

$$Y_1, Y_2, Y_3, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} Y \sim N[\mu, \sigma] \tag{6.117}$$

hvor de stokastiske populasjonsvariablene Y_i beskriver effekten for pasient nr. i:

$$Y_i \overset{\text{lign.(5.16)}}{=} \text{effekten for forsøkspasient nr. } i \tag{6.118}$$

Lign.(6.117) betyr blant annet at $E[Y] = E[Y_i] = \mu$ og at $\sigma^2[Y] = \sigma^2[Y_i] = \sigma^2$.



Figur 6.13: 95 %-konfidensintervall for σ^2.

Finn et asymptotisk konfidensintervall

$$[\, LB_n^{\sigma}\, ,\, UB_n^{\sigma},\,] \tag{6.119}$$

for den forventningsrette estimatoren $\hat{\sigma}$, hvor:

$$\hat{\sigma}^2 \;=\; S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \tag{6.120}$$

med signifikansnivå $\alpha = 0.05$.

Løsning:

Definisjonen av et asymptotisk konfidensintervall er gitt ved lign.(6.46) med $\theta = \sigma$:

$$\boxed{\lim_{n\to\infty} P(\, LB_n^{\sigma} \leq \sigma \leq UB_n^{\sigma}\,) \;=\; 1-\alpha} \qquad (6.121)$$

På samme måte som i forrige eksempel så finner man at intervallet

$$\boxed{[\; LB_n^{\sigma} \;,\; UB_n^{\sigma},\;] \;=\; \left[\; \sqrt{\frac{n-1}{(n-1)+z_{1-\alpha/2}\sqrt{2(n-1)}}}\, S_y \;,\; \sqrt{\frac{n-1}{(n-1)+z_{\alpha/2}\sqrt{2(n-1)}}}\, S_y \;\right]} \qquad (6.122)$$

er et *asymptotisk* $(1-\alpha)\,100\,\%$-konfidensintervall for σ.

Fra dataene $y_1, y_2 \ldots y_{100}$ fra forsøksrekken (se tabell 5.2 side 271) samt $s_y^2 = 0.0017$ fra tabell 5.5 side 278 får vi en *realisering* (små $y_1, y_2, \ldots, y_{100}$) av den nedre og øvre grensen: ($n = 100$)

$$\underline{LB_n^\sigma(y_1, y_2, \ldots, y_n)} = \sqrt{\frac{n-1}{(n-1) + z_{1-\alpha/2}\sqrt{2(n-1)}}}\, s_y \tag{6.123}$$

$$= \sqrt{\frac{100-1}{(100-1) + 1.96 + \sqrt{2(100-1)}}} \cdot 0.0413 \tag{6.124}$$

$$= \underline{0.0365} \tag{6.125}$$

$$\underline{UB_n^\sigma(y_1, y_2 \ldots, y_n)} = \sqrt{\frac{n-1}{(n-1) + z_{\alpha/2}\sqrt{2(n-1)}}}\, s_y \tag{6.126}$$

$$= \sqrt{\frac{100-1}{(100-1) - 1.96\sqrt{2(100-1)}}} \cdot 0.0413 \tag{6.127}$$

$$= \underline{0.0486} \tag{6.128}$$

som gir realiseringen

$$\underline{\underline{[\ LB_n^\sigma\ ,\ UB_n^\sigma\ ]}} = [\ 0.0365\ ,\ 0.0486\ ] \tag{6.129}$$

av det asymptotiske 95 %-konfidensintervallet for standardavviket σ til effekten Y.

■

6.4 Student's t-fordeling og χ^2_k-fordeling

I forrige avsnitt konstruerte vi asymptotiske 95%-konfidensintervaller for (μ, σ) for effekten av legemiddelet, se lign.(6.107) og (6.122):

$$[\, LB_n^{\mu}\,,\, UB_n^{\mu}\,] \overset{\text{lign.(6.107)}}{=} \left[\, \overline{Y} \;-\; \frac{z_{1-\alpha/2}}{\sqrt{n}}\, S_y \;,\; \overline{Y} \;-\; \frac{z_{\alpha/2}}{\sqrt{n}}\, S_y \,\right] \tag{6.130}$$

$$[\; LB_n^{\sigma},\; UB_n^{\sigma}\;] \overset{\text{lign.(6.122)}}{=} \left[\, \sqrt{\frac{n-1}{(n-1)+z_{1-\alpha/2}\sqrt{2(n-1)}}}\, S_y \;,\; \sqrt{\frac{n-1}{(n-1)+z_{\alpha/2}\sqrt{2(n-1)}}}\, S_y \,\right] \tag{6.131}$$

hvor

$$X = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{6.132}$$

$$S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 \tag{6.133}$$

$$z_{\alpha/2} = \text{nedre kvantil til } N[0,1]\text{-fordelingen} \tag{6.134}$$

$$z_{1-\alpha/2} = \text{øvre kvantil til } N[0,1]\text{-fordelingen} \tag{6.135}$$

$$n = \text{antall forsøk (antall pasienter)} \tag{6.136}$$

$$\alpha = \text{signifikansnivået} \tag{6.137}$$

Spørsmål:

Kan vi konstruere 95 %-konfidensintervaller for (μ, σ) for effekten av legemiddelet selv når antall forsøk er lavt, f.eks. $n \lesssim 30$? [16]

Sagt med andre ord:
Kan vi finne eksakte 95 %-konfidensintervaller for (μ, σ)? [17]

Svar:

Svaret er ja, men da må vi introdusere to nye sannsynlighetsfordelinger:

student's t-fordelingen for eksakt konfidensintervall for μ

χ^2_k-fordelingen for eksakt konfidensintervall for σ

[16] Grensen på 30 forsøk kommer fra gyldigheten til sentralgrenseteoremet, typisk $n \gtrsim 30$, se lign.(4.172) på side 248.

[17] Dvs. ikke asymptotiske.

6.4.1 χ^2_k-fordeling ("kji"-fordeling)

Definisjon: (χ^2_k-fordeling)

La $Z_1, Z_2, \ldots, Z_n$ være i.i.d. standard normalfordelte stokastiske variabler, dvs. $Z_i \sim N[0,1]$. Da er:

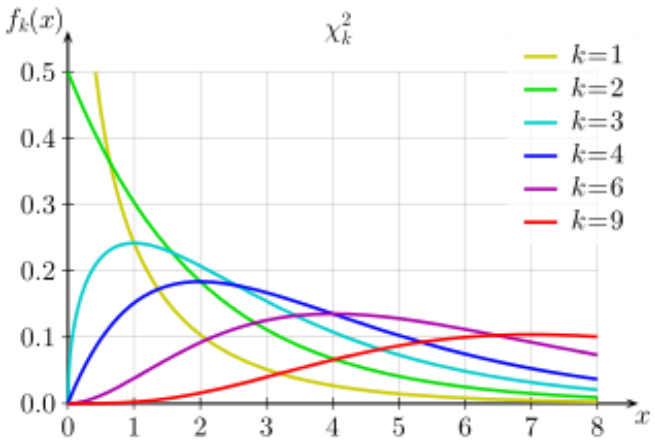
$$K = \sum_{i=1}^{n} Z_i^2 = Z_1^2 + \cdots + Z_k^2 \qquad (6.138)$$

kji-fordelt med n frihetsgrader. Vi skriver:

$$K \sim \chi^2_n \qquad (6.139)$$

■

Figur 6.14 viser χ^2_k-fordelingen for ulike ulike frihetsgrader k.



Figur 6.14: χ^2_k-fordelingen.

Setning: (S_y^2 er χ^2_{n-1}-fordelt)

La $Y_1, Y_2, \ldots, Y_n$ være i.i.d. normalfordelte stokastiske variabler, dvs. $Y_i \sim N[\mu, \sigma]$.
Da er:

$$K = \frac{S_y^2}{\sigma^2/(n-1)} \sim \chi^2_{n-1} \qquad (6.140)$$

χ^2_{n-1}-fordelt med $n-1$ frihetsgrader, hvor

$$S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 \qquad (6.141)$$

■

Kommentar

- Dette resultatet er svært viktig siden det gir oss fordelingen til estimatoren $\hat{\sigma}^2 = S_y^2$ for variansen σ^2. Vi kan dermed bruke dette teoremet for å konstruere et eksakt 95 %-konfidensintervall for σ.

Men først introduserer vi Student's t-fordelingen som bygger på kji-fordelingen og som gir oss muligheten for å konstruere et eksakt 95 %-konfidensintervall for μ.

6.4.2 Student's t-fordeling

Definisjon: (Student's t-fordeling)

La $Z \sim N[0,1]$ og $K \sim \chi^2_n$ være stokastiske variabler.
Da er:

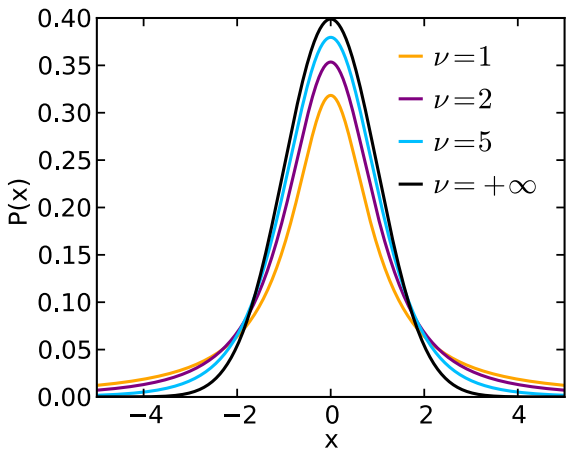
$$T = \frac{\sqrt{n}Z}{\sqrt{K}} \tag{6.142}$$

Student's t-fordelt med n frihetsgrader. Vi skriver

$$T \sim t_n \tag{6.143}$$

■

Figur 6.15 viser Student's t-fordelingen for ulike ulike frihetsgrader.



Figur 6.15: Student's t-fordelingen.

Setning: (Student's t-fordeling)

La $Y_1, Y_2, \ldots, Y_n$ være i.i.d. normalfordelte stokastiske variabler, dvs. $Y_i \sim N[\mu, \sigma]$.
Da er den stokastiske variabelen

$$\boxed{T = \frac{\overline{Y} - \mu}{S_y/\sqrt{n}} \sim t_{n-1}} \qquad (6.144)$$

Student's t-fordelt med $n-1$ frihetsgrader.

■

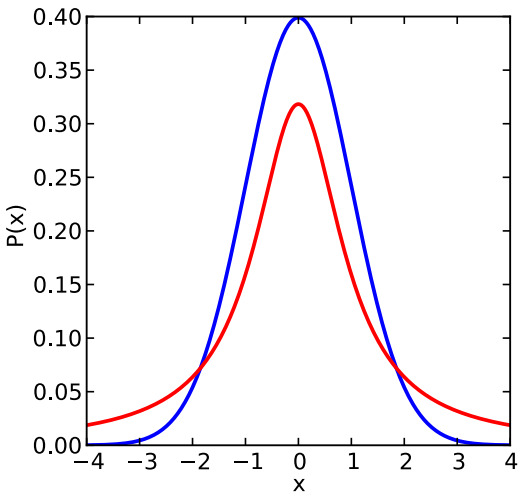
Bevis:

Vi hopper over beviset i dette kompendiet.

■

Kommentarer

- Student's t-fordelingen er svært lik normalfordelingen, men for liten n er halene mye "tykkere" enn halene til normalfordelingen, se figur 6.16.
- Når n vokser, blir Student's t-fordelingen mer og mer lik normalfordelingen, og i grensen når $n \to \infty$, er de helt like.
- Antall frihetsgrader uttrykker graden av usikkerhet som skyldes at σ er estimert.
  Stor frihetsgrad betyr liten usikkerhet.



Figur 6.16: Student's t-fordelingen (rød) med $n = 1$ frihetsgrad sammenlignet med normalfordelingen (blå).

6.4.3 Eksakte $(1-\alpha)$-konfidensintervaller for μ og σ

Eksempel: (eksakte 95%-konfidensintervall for μ av *effekten* av legemiddel)

La oss se på eksemplet med legemiddel:

$$Y \overset{\text{lign.(6.23)}}{=} \text{effekten av legemiddelet for en } \underbrace{\text{tilfeldig valgt pasient i populasjonen}}_{\text{alle som har sykdommen, ikke bare utvalget } n=100} \tag{6.145}$$

hvor $Y \sim N[\mu, \sigma]$ med:

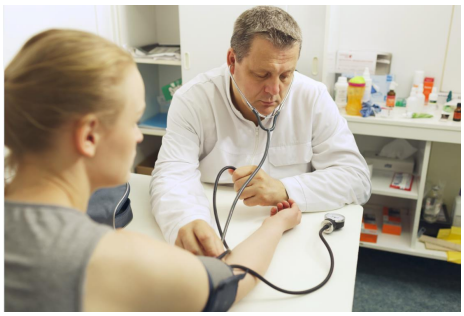
$$E[Y] = \mu \tag{6.146}$$

$$Var[Y] = \sigma^2 \tag{6.147}$$

Bruker resultatet i lign. (6.144) for å konstruere et eksakt $(1-\alpha)$-konfidensintervall for μ i eksempelet med legemiddel:

$$T = \frac{\overline{Y}-\mu}{S_y/\sqrt{n}} \sim t_{n-1} \quad \text{(Student's } t\text{-fordeling)} \tag{6.148}$$

Foto: Colourbox



Figur 6.17: Forsøk.

Løsning:

Vi bruker Student's t-fordelingens kvantiler, $q_{\alpha/2}$ og $q_{1-\alpha/2}$, for å konstruere et eksakt 95 %-konfidensintervall for μ:

$$P\Big(\; q_{\alpha/2} \; \leq \; T \; \leq \; q_{1-\alpha/2} \; \Big) \; = \; 1-\alpha \tag{6.149}$$

$$\Updownarrow \qquad \text{(algebra)}$$

$$P\Bigg(\underbrace{\overline{Y} - \frac{q_{1-\alpha/2}}{\sqrt{n}}\, S_y}_{= \, LB_n^{\mu}} \; \leq \; \mu \; \leq \; \underbrace{\overline{Y} - \frac{q_{\alpha/2}}{\sqrt{n}}\, S_y}_{= \, UB_n^{\mu}} \Bigg) \; = \; 1-\alpha \tag{6.150}$$

hvor vi definerer dermed nedre og øvre grenser:

$$LB_n^{\mu} \quad = \quad \overline{Y} - \frac{q_{1-\alpha/2}}{\sqrt{n}}\, S_y \tag{6.151}$$

$$UB_n^{\mu} \quad = \quad \overline{Y} - \frac{q_{\alpha/2}}{\sqrt{n}}\, S_y \tag{6.152}$$

hvor

$$q_{\alpha/2} \quad = \quad \text{nedre kvantil til Student's } t\text{-fordelingen} \tag{6.153}$$

$$q_{1-\alpha/2} \quad = \quad \text{øvre kvantil til Student's } t\text{-fordelingen} \tag{6.154}$$

Intervallet

$$[\ LB_n^{\mu}\ ,\ UB_n^{\mu}\ ]\ =\ \left[\ \overline{Y} - \frac{q_{1-\alpha/2}}{\sqrt{n}}\,S_y\ ,\ \overline{Y} - \frac{q_{\alpha/2}}{\sqrt{n}}\,S_y\ \right] \tag{6.155}$$

er dermed et eksakt $(1-\alpha)\,100\,\%$-konfidensintervall for μ for effekten Y.

Med $\alpha = 0.05$ og $n = 100$ for effekten Y så finner vi ved tabelloppslag: [18]

$$q_{\alpha/2} \quad = \quad q_{0.025} \ = -1.984 \tag{6.158}$$

$$q_{1-\alpha/2} \quad = \quad q_{0.975} \ = \ \ 1.984 \tag{6.159}$$

[18] Til sammenligning har vi

$$z_{\alpha/2} \quad = \quad z_{0.025} \ = -1.96 \tag{6.156}$$

$$z_{1-\alpha/2} \quad = \quad z_{0.975} \ = \ \ 1.96 \tag{6.157}$$

for N-fordeling, altså tykkere "hale" i Student's t-fordeling.

Fra dataene $y_1, y_2 \ldots y_{100}$ fra forsøksrekken (se tabell 5.2 side 271) samt $s_y = 0.0413$ fra tabell 5.5 side 278 får vi en *realisering* (små $y_1, y_2, \ldots, y_{100}$ av den nedre og øvre grensen: ($n = 100$)

$$\underline{LB_n^{\mu}(y_1, y_2, \ldots, y_n)} = \overline{y} - \frac{q_{1-\alpha/2}}{\sqrt{n}}\, s_y \tag{6.160}$$

$$= 0.8967 - \frac{1.984}{\sqrt{100}} \cdot 0.0413 \tag{6.161}$$

$$= \underline{0.8155} \tag{6.162}$$

$$\underline{UB_n^{\mu}(y_1, y_2, \ldots, y_n)} = \overline{y} - \frac{q_{\alpha/2}}{\sqrt{n}}\, s_y \tag{6.163}$$

$$= 0.8967 - \frac{(-1.984)}{\sqrt{100}} \cdot 0.0413 \tag{6.164}$$

$$= \underline{0.9444} \tag{6.165}$$

som gir realiseringen [19]

$$\underline{\underline{[\, LB_n^{\mu}\ ,\ UB_n^{\mu}\,]}} = [\, 0.8155\ ,\ 0.9444\,] \tag{6.167}$$

av det eksakte 95 %-konfidensintervallet for μ.

■

[19] Det tilsvarende asymptotiske 95 %-konfidensintervallet for μ fant i lign.(6.116) på side 333:

$$[\, LB_n^{\mu}\ ,\ UB_n^{\mu}\,] = [\, 0.8886\ ,\ 0.9048\,] \tag{6.166}$$

altså det eksakte intevallet er større enn det asymptotiske.

Eksempel: (eksakte 95%-konfidensintervall for σ av *effekten* av legemiddel)

La oss se på eksemplet med legemiddel:

$$Y \overset{\text{lign.(6.23)}}{=} \text{effekten av legemiddelet for en } \underbrace{\text{tilfeldig valgt pasient i populasjonen}}_{\text{alle som har sykdommen, ikke bare utvalget } n=100} \tag{6.168}$$

hvor $Y \sim N[\mu, \sigma]$ med:

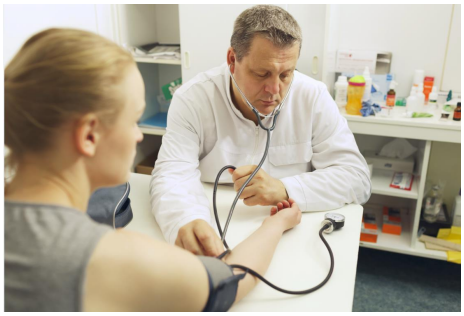
$$E[Y] = \mu \tag{6.169}$$

$$Var[Y] = \sigma^2 \tag{6.170}$$

Bruker resultatet i lign. (6.144) for å konstruere et eksakt $(1-\alpha)$-konfidensintervall for σ i eksempelet med legemiddel:

$$K = \frac{S_y^2}{\sigma^2/(n-1)} \sim \chi^2_{n-1} \tag{6.171}$$

Foto: Colourbox



Figur 6.18: Forsøk.

Løsning:

Vi bruker χ^2_{n-1}-fordelingens α kvantiler, $\chi_{\alpha/2}$ og $\chi_{1-\alpha/2}$, for å konstruere et eksakt 95 %-konfidensintervall for σ:

$$P\left(\chi_{\alpha/2} \leq K \leq \chi_{1-\alpha/2} \right) = 1-\alpha \tag{6.172}$$

$$\Updownarrow \quad \text{(algebra)}$$

$$P\left(\underbrace{\sqrt{\frac{n-1}{\chi_{1-\alpha/2}}}\, S_y}_{=\, LB_n^\sigma} \leq \sigma \leq \underbrace{\sqrt{\frac{n-1}{\chi_{\alpha/2}}}\, S_y}_{=\, UB_n^\sigma} \right) = 1-\alpha \tag{6.173}$$

hvor vi definerer dermed nedre og øvre grenser:

$$LB_n^\sigma = \sqrt{\frac{n-1}{\chi_{1-\alpha/2}}}\, S_y \tag{6.174}$$

$$UB_n^\sigma = \sqrt{\frac{n-1}{\chi_{\alpha/2}}}\, S_y \tag{6.175}$$

med

$$\chi_{\alpha/2} = \text{nedre kvantil til } \chi^2_{n-1}\text{-fordelingen} \tag{6.176}$$

$$\chi_{1-\alpha/2} = \text{øvre kvantil til } \chi^2_{n-1}\text{-fordelingen} \tag{6.177}$$

Intervallet

$$[\, LB_n^{\sigma} \,,\, UB_n^{\sigma} \,] \;=\; \left[\, \sqrt{\frac{n-1}{\chi_{1-\alpha/2}}}\, S_y \;,\; \sqrt{\frac{n-1}{\chi_{\alpha/2}}}\, S_y \,\right] \tag{6.178}$$

er dermed et eksakt $(1-\alpha)\,100\,\%$-konfidensintervall for σ for effekten Y.

Med $\alpha = 0.05$ og $n = 100$ for effekten Y så finner vi ved tabelloppslag: [20]

$$\chi_{\alpha/2} \;=\; \chi_{0.025} \;=\; 73.4 \tag{6.179}$$

$$\chi_{1-\alpha/2} \;=\; \chi_{0.975} \;=\; 128.4 \tag{6.180}$$

[20] Kvantilene $\chi_{\alpha/2}$ og $\chi_{1-\alpha/2}$ er forskjellige fordi χ-fordelingen ikke er symmetrisk, se figur (6.14) side 340.

Fra dataene $y_1, y_2 \ldots y_{100}$ fra forsøksrekken (se tabell 5.2 side 271) samt $S_y^2 = 0.0017$ fra tabell 5.5 side 278 får vi en *realisering* (små $y_1, y_2, \ldots, y_{100}$ av den nedre og øvre grensen: ($n = 100$)

$$\underline{LB_n^\sigma(y_1, y_2, \ldots, y_n)} = \sqrt{\frac{n-1}{\chi_{1-\alpha/2}}}\, s_y \tag{6.181}$$

$$= \sqrt{\frac{100-1}{128.4}} \cdot 0.0413 \tag{6.182}$$

$$= \underline{0.0362} \tag{6.183}$$

$$\underline{UB_n^\sigma(y_1, y_2, \ldots, y_n)} = \sqrt{\frac{n-1}{\chi_{\alpha/2}}}\, s_y \tag{6.184}$$

$$= \sqrt{\frac{100-1}{73.4}} \cdot 0.0413 \tag{6.185}$$

$$= \underline{0.0479} \tag{6.186}$$

som gir realiseringen [21]

$$\underline{\underline{[\ LB_n^\sigma\ ,\ UB_n^\sigma\ ]}} = [\ 0.0362\ ,\ 0.0479\ ] \tag{6.188}$$

av det eksakte 95 %-konfidensintervallet for standardavviket σ til effekten Y.

■

[21] Det tilsvarende asymptotiske 95 %-konfidensintervallet for σ fant i lign.(6.129) på side 337:

$$[\ LB_n^\sigma\ ,\ UB_n^\sigma\ ] = [\ 0.0365\ ,\ 0.0486\ ] \tag{6.187}$$

altså det er liten forskjell mellom det asymptotiske og det eksakte konfidensintervallet. Dette skyldes at $n = 100$ er stor. I grensen når $n \to \infty$ så blir de like.

Konfidensintervall for p

Asymptotisk: (CLT , normalfordeling)c

$$[\,LB_n^p\,,\,UB_n^p\,] \;=\; \left[\;\overline{X}\;-\;\frac{z_{1-\alpha/2}}{\sqrt{n}}\sqrt{\overline{X}(1-\overline{X})}\;,\;\overline{X}\;-\;\frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{\overline{X}(1-\overline{X})}\;\right]$$

= [0.8163 , 0.9437] (α = 0.05)

Konfidensintervall for μ

Asymptotisk: (CLT , normalfordeling)

$$[\,LB_n^\mu\,,\,UB_n^\mu\,] \;=\; \left[\;\overline{Y}\;-\;\frac{z_{1-\alpha/2}}{\sqrt{n}}\,S_y\;,\;\overline{Y}\;-\;\frac{z_{\alpha/2}}{\sqrt{n}}\,S_y\;\right]$$

= [0.8886 , 0.9048] (α = 0.05)

Eksakt: (Student's *t* fordeling)

$$[\;LB_n^\mu\;,\;UB_n^\mu\;] \;=\; \left[\;\overline{Y}-\frac{q_{1-\alpha/2}}{\sqrt{n}}\,S_y\;,\;\overline{Y}-\frac{q_{\alpha/2}}{\sqrt{n}}\,S_y\;\right]$$

= [0.8155 , 0.9444] (α = 0.05)

Konfidensintervall for σ

Asymptotisk: (CLT , normalfordeling)

$$[\;LB_n^\sigma\;,\;UB_n^\sigma,\;] \;=\; \left[\;\sqrt{\frac{n-1}{(n-1)+z_{1-\alpha/2}\sqrt{2(n-1)}}}\,S_y\;,\;\sqrt{\frac{n-1}{(n-1)+z_{\alpha/2}\sqrt{2(n-1)}}}\,S_y\;\right]$$

= [0.0365 , 0.0486] (α = 0.05)

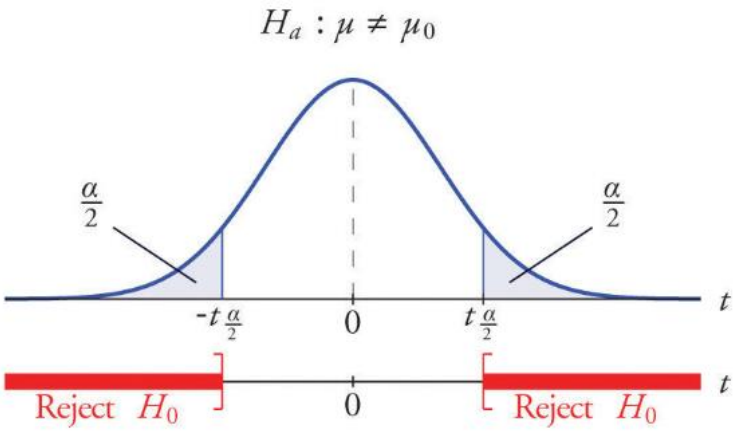
Eksakt: (χ^2 fordeling)

$$[\;LB_n^\sigma\;,\;UB_n^\sigma\;] \;=\; \left[\;\sqrt{\frac{n-1}{\chi_{1-\alpha/2}}}\,S_y\;,\;\sqrt{\frac{n-1}{\chi_{\alpha/2}}}\,S_y\;\right]$$

= [0.0362 , 0.0479] (α = 0.05)

Kapittel 7

Hypotesetesting



Figur 7.1: Hypotesetesting

7.1 Motivasjon - hypotesetesting

I legemiddelforsøket med $n = 100$ pasienter, ble den forventede effekten μ estimert til: [1]

$$\underbrace{\hat{y}(y_1, \ldots y_{100}) \;=\; \overline{y} \;=\; 0.8967}_{\text{estimat for den sanne } \mu} \tag{7.1}$$

se lign.(5.31) på side 279.

Vi husker fra side 272 at legemiddelet karakteriseres som en *suksess* dersom *effekten* av legemiddelet er 0.85 eller større:

$$\geq \; 0.85 \quad \Rightarrow \quad \text{frisk} \tag{7.2}$$
$$< \; 0.85 \quad \Rightarrow \quad \text{ikke frisk} \tag{7.3}$$

[1] Dvs. $E[Y] = \mu$, hvor μ er den sanne forventningsverdien til Y.

Spørsmål:

Finnes det nok bevis i datasettet $y_1, y_2 \dots y_{100}$ for å kunne konkludere at $\mu > 0.85$? (7.4)

Løsning:

Vi setter opp to *hypoteser*:

$$H_0 \quad : \quad \mu < 0.85 \qquad \text{(nullhypotesen)} \tag{7.5}$$

$$H_1 \quad : \quad \mu \geq 0.85 \qquad \text{(alternativ-hypotesen)} \tag{7.6}$$

Vi må velge mellom to beslutninger:

1. Forkaste H_0 og påstå H_1,
   dvs. det finnes tilstrekkelig bevis i datasettet for å konkludere at $\mu \geq 0.85$.

2. Beholde H_0,
   dvs. det finnes *ikke* tilstrekkelig bevis i datasettet for å konkludere at $\mu \geq 0.85$

Vi ønsker å designe en *funksjon* ψ, som kun er avhengig av forsøksvariablene $Y_1, \dots, Y_{100}$, som kan fortelle oss om vi kan forkaste H_0 og påstå H_1.

En slik funksjon kalles en *hypotesetest*.

7.1.1 Hypotesetest

Siden beslutningen vi skal ta er av typen ja/nei, er det naturlige å restriktere ψ til verdiene 0 og 1.

En hyptoesetest er mao. en funksjon ψ slik at:

$$\psi(Y_1, Y_2 \dots Y_{100}) \;=\; \begin{cases} 1 & , \;\; H_0 \text{ forkastes til fordel for } H_1 \\ \\ 0 & , \;\; H_0 \text{ kan } \underline{\text{ikke}} \text{ forkastes til fordel for } H_1 \text{ (beholde } H_0) \end{cases} \tag{7.7}$$

La oss lage en enkel (men kanskje ikke så god) hypotesetest:

$$\psi_1(Y_1, Y_2 \dots Y_{100}) \;=\; \begin{cases} 1 & , \;\; \overline{Y} \geq 0.85 \\ \\ 0 & , \;\; \overline{Y} < 0.85 \end{cases} \tag{7.8}$$

Fra dataene $y_1, y_2 \dots y_{100}$ fra forsøksrekken (se tabell 5.2 side 271) samt
$\overline{y} = 0.8967$ fra tabell 5.5 side 278 får vi en realisering (små $y_1, y_2, \dots, y_{100}$) av hypotesetesten:

$$\psi_1(y_1, y_2 \dots y_{100}) \;=\; 1 \qquad\qquad \text{siden } \overline{y} \;=\; 0.8967 \;\geq\; 0.85 \tag{7.9}$$

som betyr at ut fra dataene og hypotesetesten ψ_1 vil vi $\underline{\text{forkaste } H_0 \text{ til fordel for } H_1}$.

7.1.2 Tilstrekkelig bevis?

Spørsmål:

Gir ψ_1 *tilstrekkelig* bevis til å kunne forkaste H_0 og påstå H_1?
Kan vi med andre ord være relativt sikre på at konklusjonen vår er korrekt?

For å svare på dette spørsemålet, må vi studere hvilke *feile beslutninger* vi kan ta.

7.1.3 Type-I feil og type-II feil

Når vi konkluderer fra hypotesetest, kan vi gjøre *to* typer feil:

$$\text{type-I feil} \quad = \quad \text{forkaster } H_0\text{, når } H_0 \text{ er sann} \tag{7.10}$$

$$\text{type-II feil} \quad = \quad \text{beholder } H_0\text{, når } H_1 \text{ er sann} \tag{7.11}$$

For hypotesene beskrevet i lign.(7.10) og (7.11) for hvorvidt effekten til legemiddelet er lik 0.85 eller større har vi følgende tolkninger av av type-I feil og type-II feil:

$$\text{type-I feil} \quad = \quad \text{vi konkluderer at legemiddelet fungerer selv om det } \textit{ikke} \text{ fungerer} \tag{7.12}$$

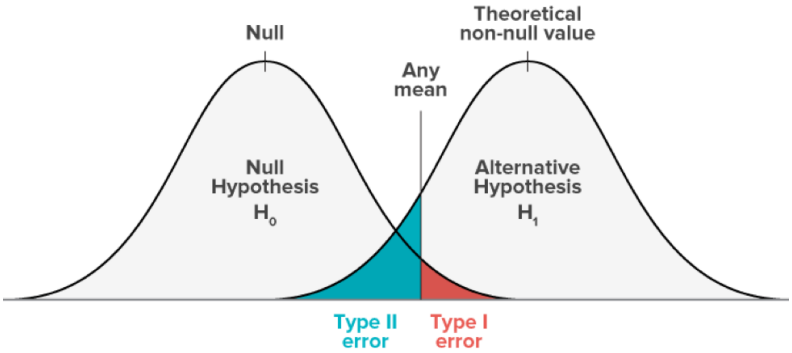
$$\text{type-II feil} \quad = \quad \text{vi konkluderer at legemiddelet } \textit{ikke} \text{ fungerer selv om det fungerer} \tag{7.13}$$

Hvilken feil synes *du* er mest alvorlig?

Tabell 7.1 viser en (klassisk) matrise som illustrerer de korrekte og feile beslutningene som kan tas ved en hypotesetest.

	H_0 **sann**	H_1 **sann**
beholde H_0	korrekt beslutning	type-II feil
forkaste H_0	type-I feil	korrekt beslutning

Tabell 7.1: Type-I og type-II feil.



Figur 7.2: Type-I og type-II feil.

7.1.4 Analogi til rettsak

Oppsettet i en hypotesetest kan sammenlignes med en rettsak hvor en person er tiltalt for drap.

$$H_0 \quad : \qquad (\text{ personen er uskyldig }) \qquad (\text{til det motsatte er bevist}) \tag{7.14}$$

$$H_1 \quad : \qquad (\text{ personen er skyldig }) \tag{7.15}$$

Så utgangspunktet i en rettsak er at personen er uskyldig (nullhypotesen) til det motsatte er bevist (alternativ-hypotesen). Dommerne kan både gjøre type-I og type-II feil: [2]

$$\text{type-I feil} \;=\; \text{dømmes selv om personen er uskyldig (justismord)} \tag{7.16}$$

$$\text{type-II feil} \;=\; \text{frifinnes selv om personen er skyldig (morder går fri)} \tag{7.17}$$

Figur 7.3: Rettsak.

[2] Legg merke til:
Dersom det ikke finnes nok bevis for å si at personen er skyldig så beholder vi H_0 (uskyldig), men det betyr ikke nødvendigvis at tiltalte var uskyldig.

Hvilken feil synes *du* er mest alvorlig?

	H_0 **sann** (uskyldig)	H_1 **sann** (skyldig)
beholde H_0 (uskyldig)	korrekt beslutning	type-II feil (morder går fri)
forkaste H_0 (skyldig)	type-I feil (justismord)	korrekt beslutning

Tabell 7.2: Type-I og type-II feil.

Det moderne rettsvesenet finner type-I feil mest alvorlig (justismord) siden dommerne skal bestemme om det foreligger tilstrekkelig bevis til å dømme personen skyldig "*utenfor enhver tvil*". Det samme gjelder hypotesetesting, men i motsetning til i en rettsak, hvor tvilen er et uklart begrep, skal vi definere eksakt hva vi mener med dette.

Begrepet

"*utenfor enhver tvil*"

kan sammenlignes med å si at

sannsynligheten for å gjøre type-I feil (justismord)

er svært lav. Men *hvor* lav?

7.1.5 Test som gir tilstrekkelig bevis

Vi sier at en hypotesetest ψ gir

$$\text{tilstrekkelig bevis} \tag{7.18}$$

i datasettet til å forkaste H_0 og godta H_1 dersom *sannsynligheten* for å gjøre type-I feil er mindre enn eller lik α: [3]

$$P_\mu(\psi_1 = 1) \;\leq\; \alpha \qquad \text{dersom } H_0 \text{ er rett, dvs. for alle } \mu < 0.85 \tag{7.19}$$

En slik test kalles en hypotesetest med *signifikansnivå* α.

[3] Typisk er $\alpha = 0.05$ eller mindre.

7.1.6 Revidert spørsmål

Gammel formulering:

Gir ψ_1 *tilstrekkelig* bevis til å kunne forkaste H_0?
Kan vi med andre ord være relativt sikre på at konklusjonen vår er korrekt?

Ny formulering:

Har hypotesetesten ψ_1 signifikansnivå α hvor $\alpha = 0.05$?

Svar:

Vi sjekker om kravet i lign.(7.19) er oppfylt for ψ_1:

$$P_{\mu=0.85}(\psi = 1) \;=\; P_{\mu=0.85}(\overline{Y} > 0.85) \;=\; \frac{1}{2} \tag{7.20}$$

siden $\overline{Y}$ er normalfordelt med forventningsverdi 0.85 dersom H_0 er rett, dvs. $\mu = 0.85$.
Mao. dersom vi konkluderer med å forkaste H_0 til fordel for H_1 via testen ψ_1, har vi en 50-50 prosent sjanse for å konkludere korrekt. Testen ψ_1 har derfor *ikke* signifikansnivå $\alpha = 0.05$.

7.1.7 En hypotesetest med signifikansnivå $\alpha = 0.05$

Vi ønsker nå å *konstruere* en hypotesetest ψ_2 som *har* signifikansnivå $\alpha = 0.05$.

Ideen er enkel:

Vi baserer testen fremdeles på gjennomsnittet $\overline{Y}$, men legger inn en ekstra *buffer* b:

$$\psi_2(Y_1, Y_2 \ldots Y_{100}) \;=\; \begin{cases} 1 & , \quad \overline{Y} \geq 0.85 + b \\ 0 & , \quad \overline{Y} < 0.85 + b \end{cases} \tag{7.21}$$

Hvor stor må bufferen b være for at testen får signifikansnivå $\alpha = 0.05$?
Vi kan forvente at bufferen er avhengig *variasjonen* til $\overline{Y}$.

Kravet i lign.(7.19) gir at:

$$P_{\mu<0.85}(\overline{Y} \geq 0.85 + b) \;\leq\; \alpha \tag{7.22}$$

Vi bruker nå setningen fra lign.(6.144) på side 343:

$$T \;=\; \frac{\overline{Y}-\mu}{S_y/\sqrt{n}} \;\sim\; t_{n-1} \tag{7.23}$$

dvs. Student's t-fordelt med $n-1$ frihetsgrader.
Standardiserer lign.(7.23):

$$P_{\mu=0.85}\big(\,\overline{Y} > 0.85 + b\,\big) \;\leq\; \alpha \tag{7.24}$$

$$P_{\mu=0.85}\left(\overbrace{\frac{\overline{Y}-0.85}{S_y/\sqrt{n}}}^{\text{Student's } t\text{-fordelt}} \;>\; \overbrace{\frac{b}{S_y/\sqrt{n}}}^{=\,q_{1-\alpha}} \right) \;\leq\; \alpha \tag{7.25}$$

hvor vi gjenkjenner uttrykket $\frac{c}{S_y/\sqrt{n}}$ i lign.(7.25) som $1-\alpha$-kvantilen $q_{1-\alpha}$ til Student's t-fordelingen.

Signifikansnivå $\alpha = 0.05$:

$$\frac{b}{S_y/\sqrt{n}} \;=\; q_{1-\alpha} \tag{7.26}$$

som gi

$$b \;=\; \frac{q_{1-\alpha}}{\sqrt{n}} S_y \tag{7.27}$$

Vi har dermed *konstruert* en hypotesetest:

$$\psi_2(Y_1, Y_2 \ldots Y_{100}) \;=\; \begin{cases} 1 & , \quad \overline{Y} \geq 0.85 \;+\; \frac{q_{1-\alpha}}{\sqrt{n}} S_y \\ 0 & , \quad \overline{Y} < 0.85 \;+\; \frac{q_{1-\alpha}}{\sqrt{n}} S_y \end{cases} \tag{7.28}$$

som har signifikansnivå $\alpha = 0.05$.

Det er vanlig å skrive en hypotesetest slik at kvantilen står alene i lign.(7.28):

$$\psi_2(Y_1, Y_2 \ldots Y_{100}) \;=\; \begin{cases} 1 & , \quad \frac{\overline{Y}-0.85}{S_y/\sqrt{n}} \geq q_{1-\alpha} \\ 0 & , \quad \frac{\overline{Y}-0.85}{S_y/\sqrt{n}} < q_{1-\alpha} \end{cases} \tag{7.29}$$

Størrelsen $T_n = \frac{\overline{Y}-0.85}{S_y/\sqrt{n}}$ kalles *teststatistikken* til hypotesetesten ψ_2.
En hypotesetest har dermed den generelle formen:

$$\psi_2(Y_1, Y_2 \ldots Y_{100}) \;=\; \begin{cases} 1 & , \quad T_n \geq c \\ 0 & , \quad T_n < c \end{cases} \tag{7.30}$$

hvor $b = q_{1-\alpha}$ i dette tilfellet. Tallet c kalles *terskelen* til hypotesetesten.

7.1.8 Hva blir konklusjonen dersom vi bruker ψ_2?

Dataene $y_1, y_2, \ldots, y_{100}$ fra forsøksrekken (se tabell 5.2 side 271) ga oss
$\overline{y} = 0.8967$ og $s_y = 0.04134$ som vi regnet ut i tabell 5.5 side 278.

Tabelloppslag: ($\overbrace{(1-\alpha)}^{\alpha\,=\,0.05}$-kvantilen $q_{1-\alpha}^{n-1}$ til Student's t-fordelingen med $\overbrace{n-1}^{n\,=\,100}$ frihetsgrader) [4]

$$q_{1-\alpha}^{n-1} \;=\; 1.645 \tag{7.31}$$

Realisering (små $y_1, y_2, \ldots, y_{100}$ av teststatistikken T_n: (til testen ψ_2)

$$T_n(y_1, \ldots, y_n) \;=\; \frac{\overline{y} - 0.85}{s_y/\sqrt{n}} \;=\; \frac{0.8967 - 0.85}{0.04134/\sqrt{100}} \;=\; 11.3075 \tag{7.32}$$

Realisering (små $y_1, y_2, \ldots, y_{100}$ av hypotesetesten: (se lign.(7.30))

$$\boxed{\psi_2(y_1, y_2 \ldots y_{100}) \overset{\text{lign.(7.30)}}{=} 1} \tag{7.33}$$

siden

$$T_n(y_1, \ldots, y_n) \;=\; 11.3075 \;>\; q_{1-\alpha}^{n-1} \;=\; 1.645 \tag{7.34}$$

som betyr at ut fra dataene og hypotesetesten ψ_1 vil vi forkaste H_0 til fordel for H_1.

[4] Til sammenligning er $z_{1-\alpha/2} = z_{1-0.05/2} = z_{0.975} = 1.65$. For store n er kvantilene $q_{1-\alpha}^{n-1}$ til Student's t-fordelingen tilnærmet lik kvantilen $z_{1-\alpha}$ til N-fordelingen.

7.2 Hypotesetesting

Definisjon (statistisk formulering - hypotesetest med signifikansnivå α)

La [5]

$$X_1, X_2 \dots X_n \overset{\text{i.i.d}}{\sim} X \ \sim \ P_\theta(X = x) \tag{7.37}$$

være en statistisk modell, hvor $\theta \in \Theta$ og verdimengde V.

La Θ_0 og Θ_1 være to *disjunkte* delmengder av parametermengden Θ,
og formuler hypotesene

$$H_0 \ : \ \theta \in \Theta_0 \qquad \text{(nullhypotesen)} \tag{7.38}$$

$$H_1 \ : \ \theta \in \Theta_1 \qquad \text{(alternativ-hypotesen)} \tag{7.39}$$

hvor målet er å finne ut om vi kan forkaste H_0 til fordel for H_1. [6]

[5] Husk at:

$$X_i \ = \ \text{forsøksvariablene} \tag{7.35}$$
$$X \ = \ \text{populasjonsvariabel} \tag{7.36}$$

[6] Legg merke til at vi skal *ikke* finne ut om H_0 er sann, *kun* bestemme om det finnes nok bevis for å forkaste H_0 til fordel for H_1.

En hypotesetest

$$\boxed{\psi(X_1, X_2 \ldots X_n) \;=\; \begin{cases} 1 & , \quad T_n > c \\ \\ 0 & , \quad T_n \leq c \end{cases}} \tag{7.40}$$

med

$$T_n \quad = \quad T_n(X_1, X_2 \ldots X_n) \;=\; \text{teststatistikk} \tag{7.41}$$

$$c \quad = \quad \text{terskel} \tag{7.42}$$

har signifikansnivå α dersom

$$\boxed{P_\theta(\psi = 1) \;\leq\; \alpha \qquad\qquad \forall \quad \theta \in \Theta_0} \tag{7.43}$$

dvs.

dersom type-I feilen til testen er mindre enn α

■

	H_0 **sann**	H_1 **sann**
beholde H_0	korrekt beslutning , $1-\alpha$	type-II feil
forkaste H_0	type-I feil , α	korrekt beslutning

Tabell 7.3: Signifikansnivå α.

Eksempel: (effekten av legemiddel - test om $\mu > 0.85$)

Bruk den generelle teorien fra tidligere i avsnittet til å sette opp en hypotesetest for effekten Y av legemiddelet.

Løsning:

Vi formulerer følgende hypoteser:

$$H_0 \quad : \quad \mu = 0.85 \qquad \text{(nullhypotesen)} \tag{7.44}$$

$$H_1 \quad : \quad \mu > 0.85 \qquad \text{(alternativ-hypotesen)} \tag{7.45}$$

Koblingen til den generelle definisjonen:

$$\theta = \mu \tag{7.46}$$

$$\Theta_0 = \{0.85\} \qquad \text{(nullhypotesen)} \tag{7.47}$$

$$\Theta_1 = (0.85, 1] \qquad \text{(alternativ-hypotesen)} \tag{7.48}$$

Figur 7.4: Forsøk.

I lign.(7.29) på side 367 en hypotesetest med signifikansnivå $\alpha = 0.05$:

$$\psi_2(Y_1, Y_2 \ldots Y_{100}) \;=\; \begin{cases} 1 & , \quad T_n > c \\ \\ 0 & , \quad T_n \leq c \end{cases} \tag{7.49}$$

med terskelen

$$c \;=\; q_{1-\alpha} \tag{7.50}$$

og teststatistikk

$$T_n \;=\; \frac{\overline{Y} - 0.85}{S_y/\sqrt{n}} \tag{7.51}$$

i dette tilfellet. Tallet c kalles *terskelen* til hypotesetesten.

Type-I feilen for denne testen oppfyller

$$P_\mu(\psi_2 = 1) \;=\; P_\mu(T_n > q_{1-\alpha}) \;\leq\; \alpha \qquad\qquad \forall \;\; \mu = 0.85 \tag{7.52}$$

■

7.3 To-utvalgs test

Anta: (eksemplet hvor $n = 100$ pasienter)

$$Y_1, Y_2, Y_3, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} Y \sim N[\mu, \sigma] \tag{7.53}$$

hvor ($i = 1, 2, 3, \ldots, 100$)

$$Y_i = \text{forsøksvariablene for effekten} \tag{7.54}$$
$$Y = \text{populasjonsvariabel for effekten} \tag{7.55}$$

Fagfolkene som gjennomførte testingen og analyserte dataene, har kommet opp med en hypotese:

Hypotese:

Legemiddelets effekt er *lavere* for menn enn kvinner.

Målsetting:

Fagfolkene ønsker lage en hypotesetest med signifikansnivå $\alpha = 0.05$ som tester om det er tilstrekkelig bevis i datasettet for å kunne konkludere at effekten av legemiddelet er *lavere* for menn enn kvinner.

Figur 7.5: Kvinner og menn.

7.3.1 Statistisk modell for to utvalg

I det tilfeldige utvalget med $n = 100$ pasienter var det $n_1 = 48$ kvinner og $n_2 = 52$ menn.
Tabell 5.2 på side 271 viser resultatene fra testene:

0.91	0.81	0.93	0.83	0.92	0.81	0.88	0.93	0.87	0.89
0.90	0.84	0.92	0.90	0.91	0.90	0.84	0.89	0.90	0.94
0.79	0.89	0.88	0.91	0.88	0.87	0.86	0.88	0.92	0.88
0.86	0.90	0.86	0.83	0.87	0.89	0.86	0.89	0.94	0.93
0.92	0.87	0.84	0.78	0.94	0.80	0.84	0.88	0.87	0.88
0.83	0.88	0.90	0.99	0.95	0.94	0.89	0.91	0.90	0.89
0.92	0.95	0.90	0.91	0.86	0.93	0.88	0.94	0.93	0.89
0.90	0.94	0.89	0.91	0.94	0.91	0.98	0.88	0.99	0.90
0.93	0.91	0.90	0.91	0.89	0.95	0.98	0.90	0.90	0.95
0.90	0.92	0.85	0.92	0.96	0.90	0.86	0.92	0.93	0.92

Tabell 7.4: Y_i - effekten.

Siden vi nå skal sammenligne kvinner og menn, deler vi datasettet i *to* grupper.
I tabell 7.5 og 7.6 har vi sortert ut kvinnene og mennene fra tabell 7.4.

0.91	0.81	0.83	0.93	0.89
0.90	0.92	0.90	0.84	0.89
0.88	0.87	0.88	0.92	0.88
0.86	0.83	0.87	0.89	0.94
0.78	0.80	0.84	0.88	0.88
0.90	0.95	0.89	0.89	0.92
0.95	0.91	0.94	0.93	0.90
0.89	0.91	0.90	0.91	0.90
0.89	0.90	0.90	0.95	0.90
0.92	0.90	0.92		

Tabell 7.5: $n_1 = 48$ kvinner.

0.93	0.92	0.81	0.88	0.87
0.84	0.90	0.91	0.89	0.90
0.94	0.79	0.88	0.91	0.86
0.90	0.86	0.89	0.86	0.93
0.92	0.87	0.84	0.94	0.87
0.88	0.83	0.99	0.94	0.91
0.90	0.90	0.86	0.93	0.88
0.89	0.94	0.91	0.94	0.98
0.88	0.99	0.93	0.91	0.95
0.98	0.92	0.85	0.96	0.86
0.93	0.92			

Tabell 7.6: $n_2 = 52$ menn.

Vi definerer nå to typer forsøksvariabler:

Kvinner:

La $Y_1 \sim N[\mu_1, \sigma_1]$ være populasjonsvariabelen for kvinner med den aktuelle sykdommen. Her er μ_1 og σ_1 ukjente parametre. Vi definerer forsøksvariablene

$$Y_{11}, Y_{12} \ldots Y_{1n_1} \overset{\text{i.i.d}}{\sim} Y_1 \sim N[\mu_1, \sigma_1] \tag{7.56}$$

hvor $n_1 = 48$ kvinner.

Menn:

La $Y_2 \sim N[\mu_2, \sigma_2]$ være populasjonsvariabelen for menn med den aktuelle sykdommen. Her er μ_2 og σ_2 ukjente parametre. Vi definerer forsøksvariablene

$$Y_{21}, Y_{22} \ldots Y_{2n_2} \overset{\text{i.i.d}}{\sim} Y_2 \sim N[\mu_2, \sigma_2] \tag{7.57}$$

hvor $n_2 = 52$ menn.

Siden det totale utvalget var tilfeldig, så kan vi anta at forsøksvariablene Y_{1i}-ene er gjensidig uavhengige av Y_{2i}-ene. Det samme er tilfelle for populasjonsvariablene Y_1 og Y_2.

7.3.2 Formulering av nullhypotese og alternativ hypotese

Vi skal teste om det er tilstrekkelig bevis i datasettet for å kunne konkludere at effekten av legemiddelet er lavere for menn enn kvinner. Utgangspunktet (dvs. nullhypotesen) er derfor at de er like, dvs. $\mu_1 = \mu_2$. Alternativ hypotesen hvor menn har lavere effekt blir dermed $\mu_1 > \mu_2$:

$$H_0 \quad : \quad \mu_1 = \mu_2 \qquad (\text{ nullhypotesen }) \tag{7.58}$$

$$H_1 \quad : \quad \mu_1 > \mu_2 \qquad (\text{ alternativ-hypotesen }) \tag{7.59}$$

Det er hensiktsmessig å reformulere hypotesene:

$$H_0 \quad : \quad \mu_1 - \mu_2 = 0 \qquad (\text{ nullhypotesen }) \tag{7.60}$$

$$H_1 \quad : \quad \mu_1 - \mu_2 > 0 \qquad (\text{ alternativ-hypotesen }) \tag{7.61}$$

siden vi vet fra lign.(4.180) på side 253 en lineærkombinasjon av N-fordelte stokastiske variabler fortsatt er N-fordelt:

$$\overline{Y}_1 - \overline{Y}_2 \;\sim\; N\left[\, \mu_1 - \mu_2 \;,\; \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \,\right] \tag{7.62}$$

hvor

$$\overline{Y}_1 \;=\; \sum_{i=1}^{n_1} Y_{1i} \tag{7.63}$$

$$\overline{Y}_2 \;=\; \sum_{i=1}^{n_2} Y_{2i} \tag{7.64}$$

7.3.3 Konstruksjon av hypotesetest med signifikansnivå $\alpha = 0.05$

Ideen bak testen er den samme som for hypotesetesten som vi konstruerte for hele utvalget. Vi konstruerer testen ψ ved å introdusere en buffer b for differansen $\overline{Y}_1 - \overline{Y}_2$:

$$\psi(Y_{11} \ldots, Y_{1n_1}, Y_{21} \ldots, Y_{2n_2}) \;=\; \begin{cases} 1 & , \quad \overline{Y}_1 - \overline{Y}_2 > 0 + b \\ 0 & , \quad \overline{Y}_1 - \overline{Y}_2 \leq 0 + b \end{cases} \tag{7.65}$$

Kravet om at type-I feilen er mindre enn eller lik $\alpha = 0.05$ kan formuleres som:

$$P\Big(\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2) > b\Big) \;=\; P\big(\overline{Y}_1 - \overline{Y}_2 > b\big) \;\leq\; \alpha \tag{7.66}$$

siden $\mu_1 - \mu_2 = 0$ når nullhypotesen er korrekt.

Standardiserer nå $\overline{Y}_1 - \overline{Y}_2$ i lign.(7.66):

$$P\left(\underbrace{\frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}}_{\underset{\sim}{\text{lign.}(4.180)}\; N} > \frac{b}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right) \quad \leq \quad \alpha \tag{7.67}$$

Siden den stokastiske variabelen i lign.(7.67) er N-fordelt, se lign.(4.180) så:

$$\frac{c}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = z_{1-\alpha} \tag{7.68}$$

hvor

$$z_{1-\alpha} = 1-\alpha\text{-kvantilen til } N[0,1]\text{-fordelingen} \tag{7.69}$$

Problemet er at σ_1 og σ_2 er ukjente, så vi må erstatte de med S_{Y_1} og S_{Y_2} hhv.

Det er mulig å vise at vi har en *tilnærmet* Student's t-fordeling dersom vi erstatter σ_1 og σ_2 med S_{y_1} og S_{y_2} i lign.(4.180): [7]

$$\frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{S_{y_1}^2}{n_1} + \frac{S_{y_2}^2}{n_2}}} \sim t_d \tag{7.70}$$

hvor

$$d = \text{antall frihetsgrader} = \frac{\frac{s_{y_1}^2}{n_1} + \frac{s_{y_2}^2}{n_2}}{\frac{\left(\frac{s_{y_1}^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_{y_2}^2}{n_2}\right)^2}{n_2 - 1}} \tag{7.71}$$

Vi erstatter $z_{1-\alpha}$-kvantilen i lign.(7.68) med Student's t-kvantilen $q_{1-\alpha}^d$ med d frihetsgrader:

$$\frac{b}{\sqrt{\frac{s_{y_1}^2}{n_1} + \frac{s_{y_2}^2}{n_2}}} = q_{1-\alpha}^d \tag{7.72}$$

som gir b:

$$b = \frac{q_{1-\alpha}^d}{\sqrt{\frac{s_{y_1}^2}{n_1} + \frac{s_{y_2}^2}{n_2}}} \tag{7.73}$$

hvor

$$q_{1-\alpha}^d = 1-\alpha\text{-kvantilen til Student's } t\text{-fordelingen med } d \text{ frihetsgrader} \tag{7.74}$$

[7] Ta denne formelen for giit.

Hypotesetesten

$$\psi(Y_{11},\ldots,Y_{1n_1},Y_{21}\ldots,Y_{2n_2}) \;=\; \begin{cases} 1 & , \quad \overline{Y}_1-\overline{Y}_2 > \frac{q^d_{1-\alpha}}{\sqrt{\frac{S^2_{y1}}{n_1}+\frac{S^2_{y2}}{n_2}}} \\ 0 & , \quad \overline{Y}_1-\overline{Y}_2 = \frac{q^d_{1-\alpha}}{\sqrt{\frac{S^2_{y1}}{n_1}+\frac{S^2_{y2}}{n_2}}} \end{cases} \tag{7.75}$$

på standardformen blir

$$\psi(Y_{11},\ldots,Y_{1n_1},Y_{21}\ldots,Y_{2n_2}) \;=\; \begin{cases} 1 & , \quad \frac{\overline{Y}_1-\overline{Y}_2}{\sqrt{\frac{S^2_{y1}}{n_1}+\frac{S^2_{y2}}{n_2}}} > q^d_{1-\alpha} \\ 0 & , \quad \frac{\overline{Y}_1-\overline{Y}_2}{\sqrt{\frac{S^2_{y1}}{n_1}+\frac{S^2_{y2}}{n_2}}} = q^d_{1-\alpha} \end{cases} \tag{7.76}$$

Hypotesetesten ψ har (asymptotisk) signifikansnivå α med test statistikk: [8]

$$T_n \;=\; \frac{\overline{Y}_1-\overline{Y}_2}{\sqrt{\frac{S^2_{y1}}{n_1}+\frac{S^2_{y2}}{n_2}}} \tag{7.77}$$

og terskel $q^d_{1-\alpha}$, hvor d er antall frihetsgrader som gitt i lign.(7.71)

[8] "Asymptotisk" betyr at vi oppnår det ønskede signifikansnivået når n blir tilstrekkelig stor.

7.3.4 Realisering og konklusjon

Fra tabell 7.5 og 7.6 har vi dataene

$$y_{1\,1}, y_{1\,2}, \ldots, y_{1\,n_1} \qquad (\text{ tabell 7.5 }) \tag{7.78}$$

$$y_{2\,1}, y_{2\,2}, \ldots, y_{2\,n_2} \qquad (\text{ tabell 7.6 }) \tag{7.79}$$

hvor $n_1 = 48$ (antall kvinner) og $n_2 = 52$ (antall menn).
Realiseringene $\overline{y}_1$, $\overline{y}_2$, s_{y_1} og s_{y_2}:

$$\overline{y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_{1i} = 0.8836 \tag{7.80}$$

$$s_{y_1}^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1} \left(y_1 - \overline{y}_1\right)^2 = 0.00186 \tag{7.81}$$

$$\overline{y}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} y_{2i} = 0.91098 \tag{7.82}$$

$$s_{y_2}^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_1} \left(y_2 - \overline{y}_{y_2}\right)^2 = 0.00158 \tag{7.83}$$

I tillegg regner vi ut antall frihetsgrader d:

$$d = \frac{\frac{s_{y_1}^2}{n_1} + \frac{s_{y_2}^2}{n_2}}{\frac{\left(\frac{s_{y_1}^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_{y_2}^2}{n_2}\right)^2}{n_2 - 1}} = 1.38 \cdot 10^6 \tag{7.84}$$

Tabelloppslag: ($d = n - 1 = 100 - 1 = 99$ og $\alpha = 0.05$)

$$q^d_{1-\alpha} \;=\; q^{99}_{1-0.05} \;=\; 1.660 \tag{7.85}$$

Realiseringen av teststatistikken T_n til testen ψ:

$$T_n(y_{11} \ldots, y_{1n_1}, y_{21} \ldots, y_{2n_2}) \;=\; \frac{\overline{y}_1 - \overline{y}_2}{\sqrt{\frac{s^2_{y1}}{n_1} + \frac{s^2_{y2}}{n_2}}} \;=\; -3.2930 \tag{7.86}$$

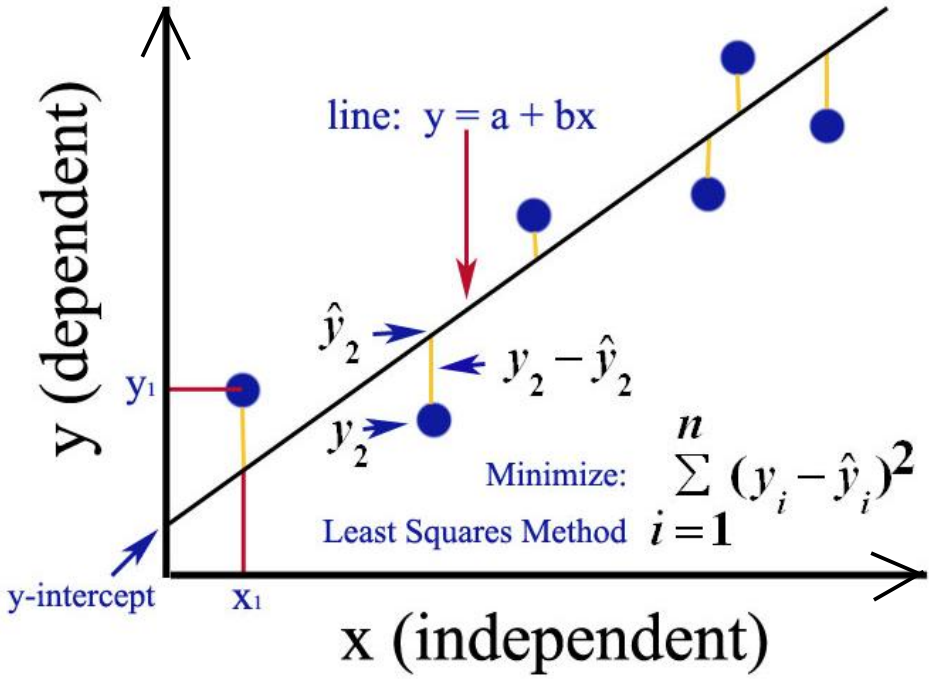
Siden

$$T_n(\, y_{11} \ldots, y_{1n_1}, y_{21} \ldots, y_{2n_2}\,) \;<\; q^d_{1-\alpha} \tag{7.87}$$

konkluderer vi fra testen og dataene at vi beholder H_0,
dvs. dataene ikke har tilstrekkelig bevis for å kunne konkludere at menn har lavere effekt på medisinen enn kvinner.

Kapittel 8

Regresjonsanalyse

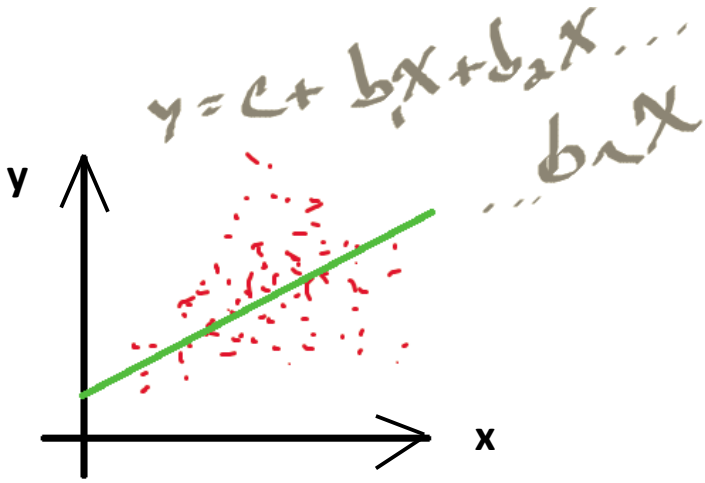


Figur 8.1: Regresjon.

8.1 Introduksjon

Regresjonsanalyse:

- Teori og metoder for å analysere og utnytte samvariasjon mellom variable.
- Formål:
  konstruere modeller som kan brukes til å anslå verdien ("prediksjon/forutsi") av en variabel Y ved hjelp av informasjon om en annen variabel X.
- terminologi:
  - variabel X: $\overbrace{\text{uavhengig variabel eller forklaringsvariabel}}^{\text{har info om dette/kjenner denne}}$
  - variabel Y: $\underbrace{\text{avhengig variabel eller responsvariabel}}_{\text{ønsker å anslå denne}}$
- Man skiller ofte mellom lineær regresjon og ikke-lineær regresjon.
- I dette kurset skal vi kun se på:
  - lineær regresjon
  - samspill mellom bare to variabler



Figur 8.2: Lineær regresjon.

8.2 Statistiske mål (to variabler)

I noen situasjoner ønsker vi å undersøke **samvariasjonen** mellom to utvalg.

Definisjon: (empirisk varians) [1]

La $x_1, x_2, x_3, ..., x_n$ være observasjoner, og la $\overline{x}$ være gjennomsnittet.
Da er den empiriske kovariansen: [2]

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \tag{8.1}$$

■

Den empiriske kovariansen s_{xy} er et mål på graden av samvariasjon mellom størrelskene x_i og tilhørende y_i.
Men den kan være vanskelig å tolke fordi:

- vi må sammenligne med andre tall som er naturlig å sammenligne med for å kunne forstå s_{xy} bedre
- s_{xy} er enhetsavhengig og gir dermed ulikt resultat dersom vi f.eks. regner med timer, minutter eller sekunder

For å gi en mer presis tolkning av graden av **SAMVARIASJON** så går vi derfor et skritt videre og definerer korrelasjonskoeffisienten r_{xy}:

[1] Kalles også **utvalgsvariansen**.
[2] Den empiriske variansen er analog til estimatoren i lign.(6.34) på side 311.

Definisjon: (korrelasjonskoeffisient)

La $x_1, x_2, x_3, ..., x_n$ og $y_1, y_2, y_3, ..., y_n$ være observasjoner. **Korrelasjonskoeffisienten** r_{xy} er da:

$$r_{xy} = \frac{s_{xy}}{s_x\, s_y} \tag{8.2}$$

■

Noen kommentarer:

- Ved å dele på s_x og s_y så får man en *normalisert* [3] versjon av s_{xy}, dvs.

$$-1 \leq r_{xy} \leq 1 \tag{8.3}$$

- r_{xy} er enhetsuavhengig
- $r_{xy} = -1$:
  - perfekt negativ korrelasjon, dvs. store x hører sammen med små y.
  - lineær[4] sammenheng mellom x og y
- $r_{xy} = 1$:
  - perfekt positiv korrelasjon, dvs. store x hører sammen med store y.
  - lineær sammenheng mellom x og y
- $r_{xy} = 0$:
  - ingen korrelasjon
  - ukorrelert
- r_{xy} er et mål på lineær korrelasjon

[3] Legg merke til begrepet *normalisert*. Det skal vi komme tilbake til ved flere anledninger.

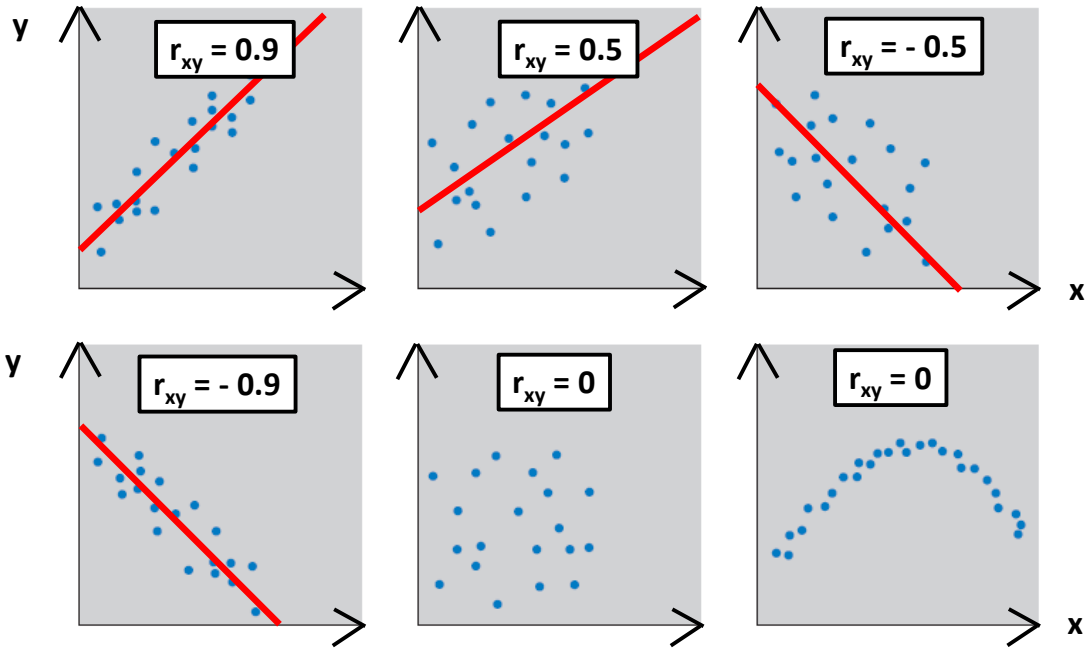
[4] Lineær sammenhenger mellom x og y betyr at de kan skrives på formen: $y = ax + b$, (a og b er konstanter). Lineær er altså det samme som en rett linje.

Eksempel: (lineær sammenheng)

La oss se nærmere på to størrelser x og y. Disse størrelsene kan være hva som helst, f.eks. pris på aksje x og pris på aksje y. Anta at disse størrelsene varierer med tiden. Anta videre at man måler x og y over en periode på 20 dager. For dag 1 er verdiene x_1 og y_1. For dag 2 har verdiene endret seg til x_2 og y_2 osv. Helt frem til dag 20 hvor størrelsene har verdiene x_{20} og y_{20}. Vi har altså samhørende **observasjoner** av par (x_i, y_i):

$$(x_1, y_1) \quad , \quad (x_2, y_2) \quad , \quad \ldots \quad , \quad (x_{19}, y_{19}) \quad , \quad (x_{20}, y_{20}) \tag{8.4}$$

La oss se på 6 forskjellige datasett som vist i figur 8.3:



Figur 8.3: Sammenheng mellom x og y samt tilhørende korrelasjonskoeffisienten r_{xy}.

Husk at r_{xy} er et mål på **lineær** sammenheng mellom x og y. For de tilfellene hvor r_{xy} er "nære" $+1$ eller -1 så er det "nære" en lineær sammenheng mellom x og y. Lineær regresjonsanalyse går ut på å:

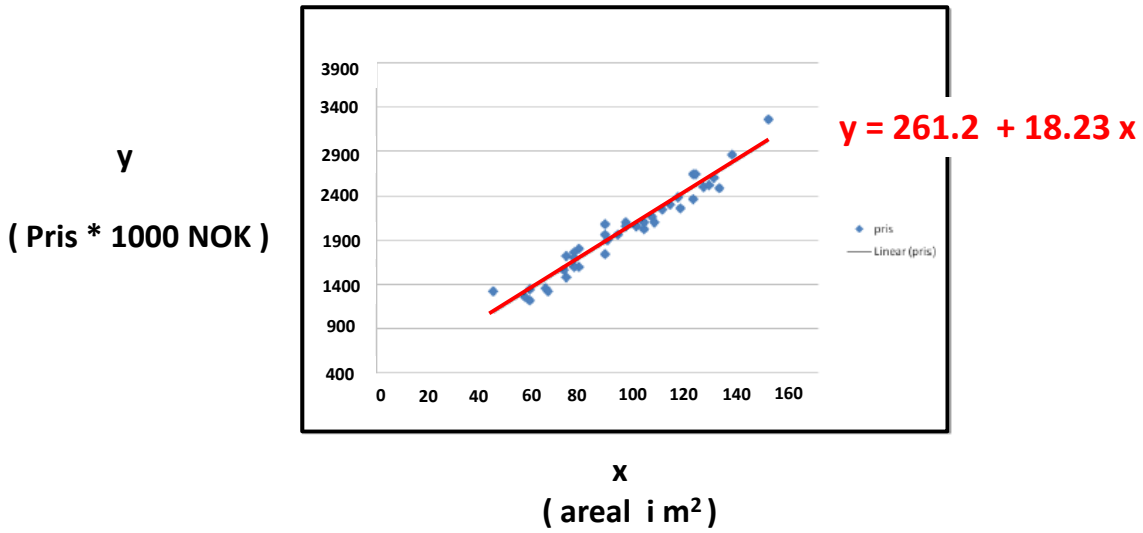
finne estimat for den rette linjen som "passer best" med datasettet

Hvorfor gir det ikke mening å finne linjen som "passer best" i de to grafene nederst til høyre i figur (8.3)?

■

Eksempel: (lineær regresjon)

La se på samvariasjon mellom utsalgspris og boareal for leiligheter i Molde. Et datasett med priser y og areal x gir spredningsplottet i figur 8.4. Vi ser en **tydelig lineær** sammenheng mellom variablene.



Figur 8.4: Sammenheng mellom y (pris) og x (areal), en regresjonslinje.

Den linjen som "passer best" kalles regresjonslinjen. For dataene i figur (8.4) er det:

$$y = 261.2 \; + \; 18.23\, x \qquad (8.5)$$

hvor y er pris på leilighet oppgitt i antall 1000 NOK. For en leilighet med boareal $x = 100$ m^2 vil den estimerte modellen predikere prisen:

$$\underline{\underline{y(100)}} \; = \; 261.2 \; + \; 18.23 \cdot 100 \; = \; \underline{\underline{20842}} \qquad (8.6)$$

dvs. litt under 2.1 mill.

Tallet 2.1 mill. er prediksjonen fra modellen for prisen på en 100 m^2 leilighet.

■

To spørsmål:

1) Hvordan finner man regresjonslinjen?

2) Siden modellen bare bruker arealet og ingen annen informasjon må vi vente noe feilmarginer. I hvor stor grad forklarer arealet x prisen y?

Disse spørsmålene skal vi besvare.

8.3 Teoretisk modell vs estimert modell

Det er viktig å skille mellom teoretisk modell og estimert modell (regresjonslinje).

Teoretisk modell:

En teoretisk modell beskriver hvordan vi tenker oss den virkelige sammenhengen mellom variablene, typisk:

$$\underbrace{y = a + bx + e}_{\text{eksakt}} \tag{8.7}$$

hvor variabelen e beskriver "avviket"[5] fra den eksakte lineære funksjonen.

Estimert modell: ($\overbrace{\text{med "hatt"}}^{\text{passer best}}$)

Parametrene a og b i lign.(8.7) er ukjente.
Men de kan estimeres ut fra et datasett med samhørende observasjoner av par:

$$\underbrace{(x_1, y_2)\ ,\ (x_2, y_2)\ ,\ \ldots\ ,\ (x_n, y_n)}_{\text{observasjoner}}\ . \tag{8.8}$$

Vi ønsker å bestemme estimater for a og b.
Disse estimatene har en "hatt" på seg, $\hat{a}$ og $\hat{b}$.
Linjen med estimatene $\hat{a}$ og $\hat{b}$ kalles **regresjonslinjen**:

$$\underbrace{\hat{y} = \hat{a} + \hat{b}x}_{\text{regresjonslinje (passer best)}} \tag{8.9}$$

Estimatene $\hat{a}$ og $\hat{b}$ bestemmes ved å finne den linje som "passer best" med datasettet.

[5] "e" står for "error".

8.4 Residual og *sse*

1) Observasjoner:

De røde punktene i figur 8.5 viser de fem observasjonspunktene (x_1, y_1), (x_2, y_2), ... , (x_5, y_5).

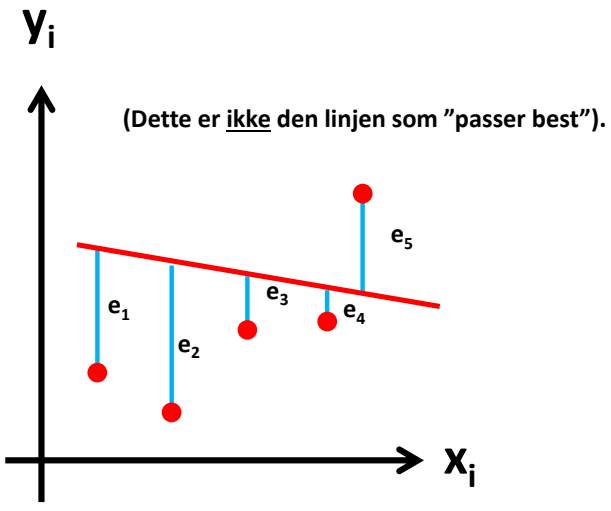
2) Rett linje:

Den røde linjen i figur 8.5 er en linje som ikke passer best til observasjonene x_i
og tilhørende verdi y_i.

Forskjellen mellom de observerte verdiene y_i og de tilsvarende verdiene til den rette linjen er de loddrette avstandene (blå linjer) som vist i figur 8.5. Forskjellen/avviket mellom observert verdi og prediksjonen som den rette linjen foreslår for datapunktet er:

$$\underbrace{e_i \;=\; y_i - \hat{y}_i}_{\text{residual}} \qquad (8.10)$$

og kalles residual eller estimat for eksperimentfeilen.
Residualen e_i måler dermed feilen vi gjør ved å bruke verdien på den rette linjen istedet for de observerte verdiene. Residualen e_i kan være positiv, negativ eller 0. [6]



Figur 8.5: Residual.

[6] Ingen residual i figur 8.5 er null. Alle er negative bortsett fra e_5.

Residualen $e_i = y_i - \hat{y}_i$:

- e_i kan være positiv, negativ eller 0
- e_i = avvik mellom estimert linje og observerte y_i-verdier.
- e_i = residual nr. i

Definisjon: (sse) [7]

La $(x_1, y_2), (x_2, y_2), ..., (x_n, y_n)$ være observasjonspar/datasett.
Størrelsen *sse*, "*sum squared error*", er da definert ved: [8]

$$sse = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{8.12}$$

hvor [9]

$$\hat{y}_i = \text{estimat/prediksjon for } y_i \tag{8.14}$$

$$y_i = \text{faktiske observasjonene/dataene nr. } i \tag{8.15}$$

■

[7] *sse* står for sum square error.

[8] Siden $e_i \overset{\text{lign.(8.10)}}{=} y_i - \hat{y}_i$ så kan *sse* alternativt skrives:

$$sse = \sum_{i=1}^{n} e_i^2 \tag{8.11}$$

[9] I vårt kurs dreier prediksjonene gitt som linjen

$$\underbrace{\hat{y}_i = \hat{a} + \hat{b}x_i}_{\text{prediksjoner}} \tag{8.13}$$

hvor $\hat{a}$ og $\hat{b}$ er det estimatene som gir en linje som "passer best" med observasjonene. Vi skal finne uttrykk for disse optimale $\hat{a}$ og $\hat{b}$ i neste avsnitt.

8.5 Minste kvadraters regresjonslinje

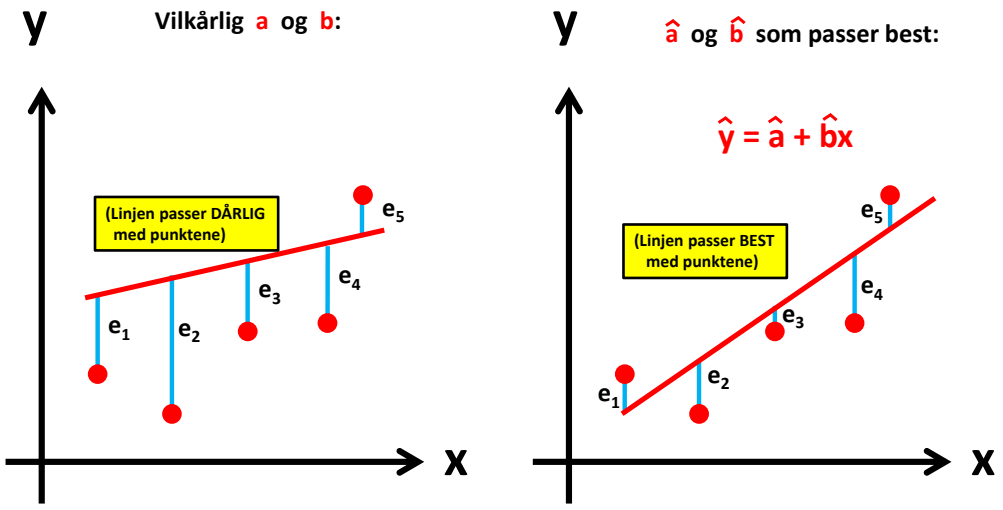
Linjen på venstre side i figur 8.6 er åpenbart ikke den som "passer best".
Linjen på høyre, derimot, passer bedre. Vi ønsker nå å finne den linjen som "passer best". Med det mener vi:

Passer best:	*sse* er minst mulig.

Å finne den linjen som passer best med datasettet er det samme som å finne den a og b som gir minst SSE, dvs. minst "*sum squared error*" i forhold til datasettet. SSE er minst der hvor stigningen er null, dvs. den deriverte med hensyn på de respektive parametrene, er lik null: [10] [11]

$$\frac{\partial sse}{\partial a} = \frac{\partial}{\partial a}\left(\sum_{i=1}^{n}\Big(y_i-(a+bx_i)\Big)^2\right) = 2(-1)\sum_{i=1}^{n}\Big(y_i-a-bx_i\Big) = 0 \qquad (8.16)$$

$$\frac{\partial sse}{\partial b} = \frac{\partial}{\partial b}\left(\sum_{i=1}^{n}\Big(y_i-(a+bx_i)\Big)^2\right) = 2(-1)\sum_{i=1}^{n}\Big(y_i-a-bx_i\Big)x_i = 0 \qquad (8.17)$$



Figur 8.6: Linje som ikke passer best, og linje som passer best.

[10] Bruker kjerneregelen som vi lærte om i "*MAT100 Matematikk*".

[11] At lign.(8.16) er et minimum, og ikke et maksimun, innser man siden $\frac{\partial^2 SSE}{\partial a^2} > 0$ og $\frac{\partial^2 SSE}{\partial a^2} > 0$.

De spesielle verdiene for a og b som minimerer SSE har fått egen notasjon, $\hat{a}$ og $\hat{b}$.
Disse er definert ved lign.(8.16). Eksplisitte uttrykk for disse finnes ved å løse nevnte ligning:

Siden gjennomsnittet $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ gir $n\overline{x} = \sum_{i=1}^{n} x_i$, og tilsvarende for y, så fås:

$$\overbrace{\sum_{i=1}^{n} y_i}^{=n\overline{y}} - \hat{a}\overbrace{\sum_{i=1}^{n}}^{=n} - \hat{b}\overbrace{\sum_{i=1}^{n} x_i}^{=n\overline{x}} = 0 \tag{8.18}$$

$$\sum_{i=1}^{n} x_i y_i - \hat{a}\underbrace{\sum_{i=1}^{n} x_i}_{=n\overline{x}} - \hat{b}\sum_{i=1}^{n} x_i^2 = 0 \tag{8.19}$$

og

$$\cancel{n}\overline{y} - \hat{a}\cancel{n} - \hat{b}\cancel{n}\overline{x} = 0 \tag{8.20}$$

$$\sum_{i=1}^{n} x_i y_i - \hat{a}\, n\, \overline{x} - \hat{b}\sum_{i=1}^{n} x_i^2 = 0 \tag{8.21}$$

Løser med hensyn på $\hat{a}$ og $\hat{b}$ alene gir:

$$\hat{a} = \overline{y} - \hat{b}\,\overline{x} \tag{8.22}$$

$$\hat{b} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \tag{8.23}$$

Disse ligningene kan skrives ved hjelp av den empiriske variansen, lign.(5.23), og den empiriske kovariansen, lign.(8.1) på følgende måte: [12]

$$\hat{a} = \overline{y} - \hat{b}\,\overline{x} \tag{8.24}$$

$$\hat{b} = \frac{s_{xy}}{s_x^2} \tag{8.25}$$

hvor

$$s_x^2 \overset{\text{lign.(5.23)}}{=} \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{8.26}$$

$$s_{xy} \overset{\text{lign.(8.1)}}{=} \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \tag{8.27}$$

[12] Man må utføre et par linjer med algebra for å innse at lign.(8.24) gir lign.(8.26). Detaljene er ikke tatt med her. Men kanskje du greier seg selv?

Setning: (minste kvadraters sum - lineære regresjonslinje)

La $(x_1, y_2), (x_2, y_2), ..., (x_n, y_n)$ være observasjonspar/datasett.
Minste kvadraters sum gir den lineære regresjonslinjen: q[13]

$$\boxed{\hat{y} = \hat{a} + \hat{b}\,x\,,} \tag{8.28}$$

hvor

$$\boxed{\begin{aligned} \hat{b} &= \frac{s_{xy}}{s_x^2} && (8.29)\\ \hat{a} &= \overline{y} - \hat{b}\,\overline{x} && (8.30) \end{aligned}}$$

og

$$s_x^2 \overset{\text{lign.(5.23)}}{=} \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{8.31}$$

$$s_{xy} \overset{\text{lign.(8.1)}}{=} \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \tag{8.32}$$

og hvor $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ og $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

■

[13] Den linjen som passer best, dvs. minst *sse*, har altså fått navnet regresjonslinje.

Eksempel: (transport - s_{xy} , regresjonslinje)

Et transportfirma har et varemottak for vogntog med spesialgods. Det tar svært lang tid å laste av et vogntog med denne type last. Transportfirmaet gjør derfor én måling per dag i en periode på 10 dager: når et tilfeldig vogntog ankommer varemottaket så teller de antall vogntog x som står foran i kø. I tillegg så måler de ventetiden y for det nylig ankomne vogntoget.

La oss definere følgende variabler:

$$x = \text{antall vogntog foran i kø} \tag{8.33}$$
$$y = \text{antall timer i ventetid} \tag{8.34}$$

For enkelhetsskyld så måler de y kun i hele timer. Resultatet er:

x_i (antall vogntog foran i kø)	2	12	1	1	10	25	3	9	27	2
y_i (antall timer ventetid)	3	11	3	1	12	21	6	4	31	2

Figur 8.7: Observasjoner x_i og y_i, hvor $i = 1, 2, ..., 10$.

Figur 8.8: Vogntog og varemottak.

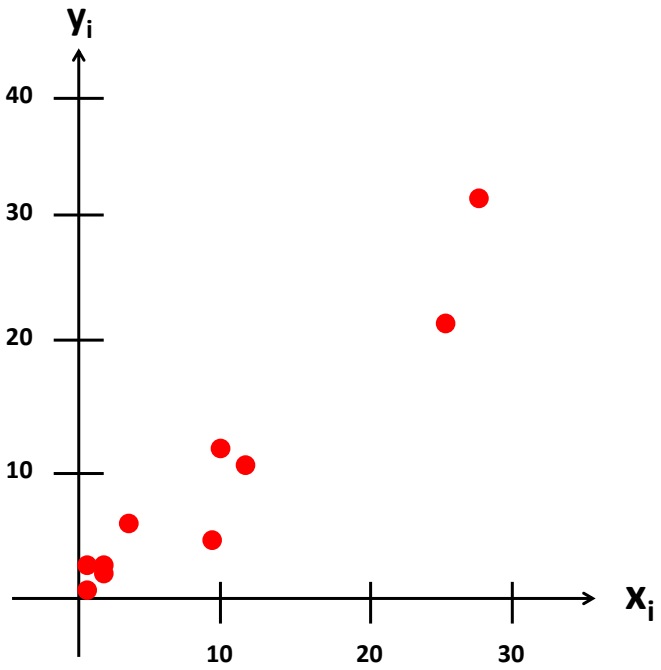
a) Gi en grafisk fremstilling av x_i og y_i. Kommenter svaret.

b) Finn gjennomsnittsverdiene $\overline{x}$ og $\overline{y}$.

c) Finn kovariansen s_{xy}.

d) Finn minste kvadraters lineæregresjonslinje for x og y.

a) Grafisk fremstilling av x_i og y_i:



Figur 8.9: Grafisk fremstilling av x_i og y_i, hvor $i = 1, 2, ..., 10$.

Kommentar:

Fra denne grafen ser vi at:

- små x_i-verdier faller sammen med små y_i-verdier
- store x_i-verdier faller sammen med store y_i-verdier

Dermed innser vi at det er en viss grad av **samsvar** mellom x_i og y_i.

b) Gjennomsnittsverdiene $\overline{x}$ og $\overline{y}$ er: ($n = 10$)

$$\underline{\underline{\overline{x}}} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{1}{10}\Big(2 + 12 + 1 + ... + 2\Big) = \underline{\underline{9.2}} \quad \text{(antall vogntog i kø)} \tag{8.35}$$

$$\underline{\underline{\overline{y}}} = \frac{1}{10}\sum_{i=1}^{10} y_i = \frac{1}{10}\Big(3 + 11 + 3 + ... + 2\Big) = \underline{\underline{9.4}} \quad \text{(ventetid, i timer)} \tag{8.36}$$

c) Vi bruker gjennomsnittsverdiene $\overline{x}$ og $\overline{y}$ når vi skal finne **kovariansen** s_{xy}: ($n = 10$)

$$\underline{\underline{s_{xy}}} \overset{\text{lign.(8.1)}}{\underset{\text{n=10}}{=}} \frac{1}{10-1}\sum_{i=1}^{10}(x_i - \overline{x})(y_i - \overline{y}) \tag{8.37}$$

$$= \frac{1}{10-1}\Big[(2 - 9.2)(3 - 9.4) + (12 - 9.2)(11 - 9.4) + ... + (2 - 9.2)(2 - 9.4)\Big]$$

$$= \underline{\underline{90.8}} \tag{8.38}$$

d) Den empiriske variansen s_x^2 kan regnes ut via definisjonen i lign.(5.23):

$$s_x^2 \overset{\text{lign.(5.23)}}{\approx} 94.6 \tag{8.39}$$

I oppgave **c** regnet vi ut kovariansen, se lign.(8.38):

$$s_{xy} \overset{\text{lign.(8.38)}}{=} 90.8 \tag{8.40}$$

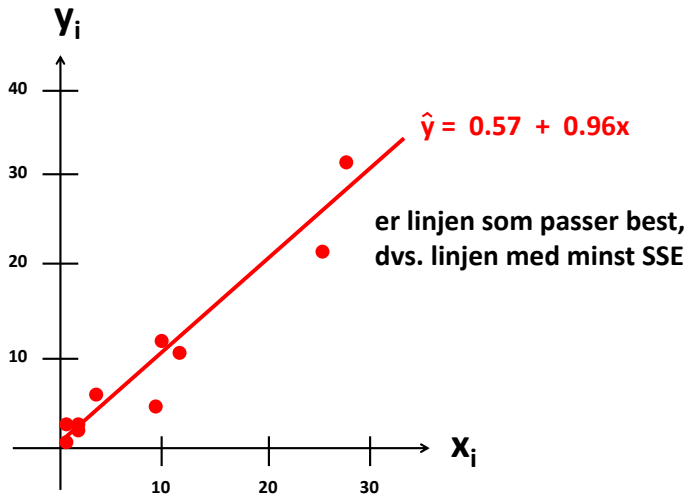
Parametrene $\hat{b}$ og $\hat{a}$ er da:

$$\underline{\hat{b}} \;=\; \frac{s_{xy}}{s_x^2} \;=\; \frac{90.8}{94.6} \;\approx\; \underline{0.96} \tag{8.41}$$

$$\underline{\hat{a}} \;=\; \overline{y} \;-\; b\overline{x} \;=\; 9.4 - 0.96 \cdot 9.2 \;\approx\; \underline{0.57} \tag{8.42}$$

Minste kvadraters lineære regresjonslinje blir dermed ifølge lign.(8.28):

$$\underline{\underline{\hat{y} \;=\; 0.57 \;+\; 0.96\, x}} \tag{8.43}$$



Figur 8.10: Grafisk fremstilling av x_i og y_i.

8.6 Forklaringsfraft og *sst*

På side 386, lærte vi at "problemet" med kovariansen s_{xy} er at den kan være vanskelig å tolke. Dette bl.a. fordi:

- størrelsen s_{xy} kan gi "store" eller "små" tall som vi må sammenligne med andre tall for å kunne forstå bedre
- s_{xy} er enhetsavhengig og gir dermed ulikt resultat dersom vi f.eks. regner med timer, minutter eller sekunder

Dette problemet ble løst ved å introdusere korrelasjonskoeffisienten r_{xy}. Denne koeffisienten har egenskaper som oppsummert på side 386.

Samme "problem" har *sse* i lign. (8.12). Hva betyr det at *sse* er "liten"? Eller "stor"? I forhold til hva? Vi løser dette problemet ved å introdusere "forklaringskraften" r^2. Før vi definerer r^2 må vi først definere en ny størrelse, nemlig *sst*:

Definisjon: (sst) [14]

La $(x_1, y_2), (x_2, y_2), ..., (x_n, y_n)$ være observasjonspar/datasett.
Størrelsen sst, "*sum squared total*", er da definert ved:

$$\boxed{sst \;=\; \sum_{i=1}^{n} (y_i - \overline{y})^2} \tag{8.44}$$

$$y_i \quad = \quad \text{faktiske observasjonene/dataene nr. } i \tag{8.45}$$

$$\overline{y} \overset{\text{lign.(5.21)}}{=} \frac{1}{n} \sum_{i=1}^{n} y_i \tag{8.46}$$

■

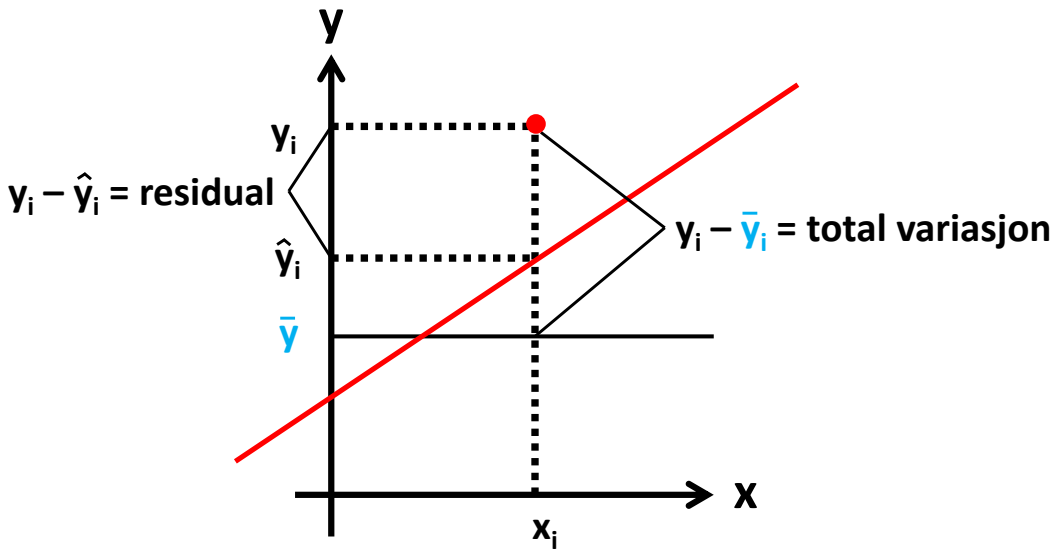
Sammenlign definisjonen ovenfor med sse fra lign.(8.12):

$$sse \;=\; \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{8.47}$$

Ser du forskjellen på lign.(8.47) og (8.44)?

[14] sst står for sum square total.

Visualisering av residual og total variasjon:



Figur 8.11: Residual og total variasjon.

Definisjon: (forklaringskraft)[15]

La $x_1, x_2, x_3, ..., x_n$ og $y_1, y_2, y_3, ..., y_n$ være observasjoner.
Forklaringskraft r^2 er da:

$$\boxed{r^2 = 1 - \frac{sse}{sst}} \tag{8.48}$$

hvor

$$sse = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{8.49}$$

$$sst = \sum_{i=1}^{n} (y_i - \overline{y})^2 \tag{8.50}$$

og

$$y_i = \text{faktiske observasjonene/dataene nr. } i \tag{8.51}$$

$$\hat{y}_i = \text{estimat/prediksjon for } y_i \tag{8.52}$$

$$\overline{y} \overset{\text{lign.(5.21)}}{=} \frac{1}{n} \sum_{i=1}^{n} y_i \tag{8.53}$$

■

[15] Kallse også forklarings*grad* eller forklarings*styrke*.

Noen kommentarer: [16]

- Forklaringskraften r^2 er *normalisert*:

$$\boxed{0 \leq r^2 \leq 1} \qquad (8.54)$$

- r^2 er enhetsuavhengig
- $r^2 = 1$:
  - alle observerte punkter ligger på regresjonslinjen
- $r^2 = 0$:
  - verdien for x har ingen betydning for verdien av y
- r^2 sier noe om:
  - hvor stor andel av den totale vasiasjonen som forklares av regresjonslinjen

Forklaringskraft r^2 og korrelasjonskoeffisienten r_{xy} som vi lærte om på side 386 har analoge egenskaper.

[16] r_{xy} og r^2 har ikke noe med hverandre å gjøre. Likevel kan det være hensiktsmessig å sammenligne egenskapene til r^2 som oppsummert her, med egenskapene til r_{xy} på side 386.

Eksempel: (logistikk - forklaringskraft)

Finn forklaringskraften r^2 for eksemplet fra side 398.

Løsning: (forklaringskraft , logistikk)

La oss se atter en gang se på eksemplet fra side 398.

x_i (antall vogntog foran i kø)	2	12	1	1	10	25	3	9	27	2
y_i (antall timer ventetid)	3	11	3	1	12	21	6	4	31	2

Figur 8.12: Samsvarende verdier av antall vogntog foran i kø x og antall timer ventetid y.

Fra lign.(8.43) vet vi:

$$\hat{y}_i \;=\; 0.57 \;+\; 0.96\, x_i \tag{8.55}$$

De faktiske verdiene y_i er vet tabellen i figur (8.12). Dermed kan regne ut sse fra lign. (8.12):

$$\begin{aligned}
\underline{sse} \quad &= \quad \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= \quad \sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}x_i)^2 \\
&= \quad (3 - 0.57 - 0.96 \cdot 2)^2 \quad + (11 - 0.57 - 0.96 \cdot 12)^2 \\
&+ \quad (3 - 0.57 - 0.96 \cdot 1)^2 \quad + (1 - 0.57 - 0.96 \cdot 1)^2 \\
&+ \quad (12 - 0.57 - 0.96 \cdot 10)^2 + (21 - 0.57 - 0.96 \cdot 25)^2 \\
&+ \quad (6 - 0.57 - 0.96 \cdot 3)^2 \quad + (4 - 0.57 - 0.96 \cdot 9)^2 \\
&+ \quad (31 - 0.57 - 0.96 \cdot 27)^2 + (2 - 0.57 - 0.96 \cdot 2)^2 \\
&\approx \quad \underline{74.2}
\end{aligned} \tag{8.56}$$

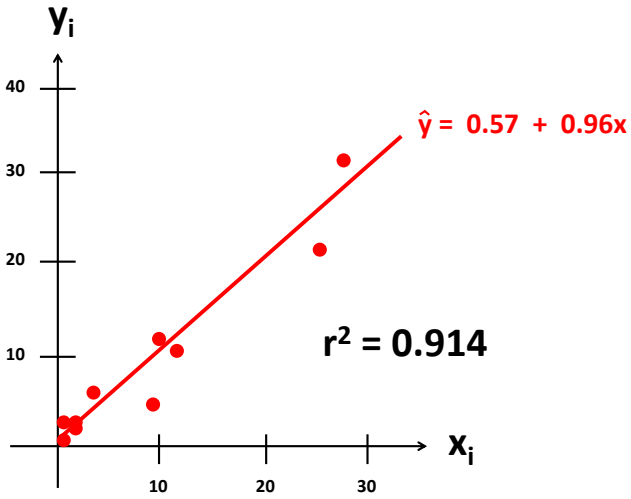
Fra lign.(8.36) vet vi gjennomsnittet $\overline{y} = 9.4$.
Dermed kan regne ut sse fra lign. (8.44):

$$\begin{aligned} \underline{sst} &= \sum_{i=1}^{n} (y_i - \overline{y})^2 \\ &= (3-9.4)^2 + (11-9.4)^2 \\ &+ (3-9.4)^2 + (1-9.4)^2 \\ &+ (12-9.4)^2 + (21-9.4)^2 \\ &+ (6-9.4)^2 + (4-9.4)^2 \\ &+ (31-9.4)^2 + (2-9.4)^2 \\ &= \underline{858.4} \end{aligned} \tag{8.57}$$

Forklaringskraft R^2 er da: (se lign.(8.48)

$$\underline{\underline{r^2}} = 1 - \frac{sse}{sst} = 1 - \frac{74.2}{858.4} = \underline{\underline{0.914}} \tag{8.58}$$

dvs. regresjonslinjen forklarer hele 91.4 % av den totale variasjonen.
Vi sier at modellen har stor forklaringskraft.

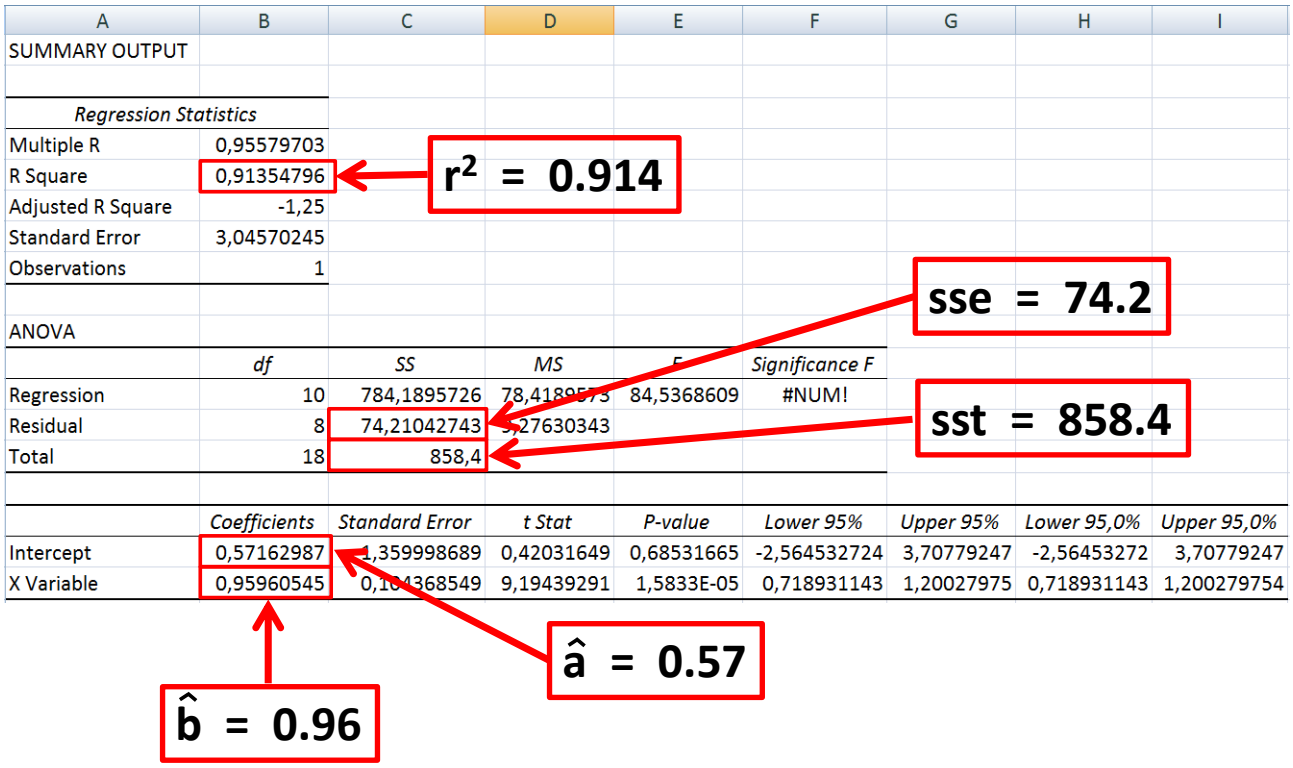


Figur 8.13: Grafisk fremstilling av x og y.

Kommentar:

Som regel gjør man ikke all denne regningen "for hånd".
Mange dataprogrammer kan hjelpe oss å regne ut statistiske størrelser, f.eks. Excel.

Dersom vi bruker Excel på eksemplet vårt så får vi en utskrift som vist i figuren nedenfor. Da kan vi lese av de størrelsene vi regnet ut "for hånd" direkte fra Excel-utskriften.



Figur 8.14: Utskrift fra Excel.

■