Statistics for Engineering

Thobias Høivik

Fall 2025

Contents

1 Introduction

This course covers introductory statistics and is taught at the Western Norway University of Applied Science. This will be among my most scattered and improvised notes, probably mostly in norwegian.

2 Lecture 1

Mål:

- Finne matematiske størrelser som beskriver data i eit datasett
- Sentralmål
- Spredningsmål

2.1 Sentralmål

- Finne ein verdi som representerer ein "typisk" enhet i ei mengde.

Example 2.1

Lønna til 13 personer oppgitt som: $110, 125, 125, 300, 350, 370, 375, 380, 390, 410, 430, 435, 440$.

Modus/typetal.

Verdien som dukkar opp flest gangar. Fra eksempellet har at vi moduslønn er 125.

Median. Median er den midterste verdien i ei datamengde. I dette tilfellet fra Example 2.1 har vi at medianlønn ligger på 375. I tilfellet hvor det er et partall antall verdier tar vi gjennomsnittet av de to midterste verdiene.

Gjennomsnitt.

Gjennomsnittet av en datamengde X er gitt ved

$$\overline{X} = \frac{1}{|X|} \sum_{i=1}^{|X|} x_i$$

I dette tilfellet, fra Example 2.1, har vi gjennomsnitt på 326.

Anta nå at en 14-ende tjener blir lagt til i mengden som tjener $30,000$. Da er gjennomsnittslønnen på de 14 personene rundt 2445 som ligger langt over den nest høgste tjeneren. Med andre ord drar person nummer 14 gjennomsnittet opp vanvittigt.

2.2 Spredningsmål.

Målet med et spredningsmål er å beskrive hvor stor variasjon det er i datasettet.

Example 2.2

Gitt to mengder X, Y.
$X = \{80, 90, 100, 110, 120\}$
$\overline{X} = 100$
$Y = \{20, 60, 100, 140, 180\}$
$\overline{Y} = 100$

Variasjonsbredde.

Variasjonsbredda er største verdi minus minste verdi:

$$\max X - \min X$$

Fra Example 2.2 har vi at at variasjonsbredda til X er 40 og variasjonsbredda til Y er 160.

Varians og standardavik.

Varians:

$$\text{var}(X) = s_X^2 = \frac{1}{|X|-1} \sum_{i=1}^{|X|} \left(x_i - \overline{X}\right)^2$$

Standardavik:

$$\text{std}(X) = \sqrt{s_X^2} = \sqrt{\text{var}(X)}$$

Viss vi ser på Example 2.3 får vi

$$\text{var}(X) = \frac{(80-100)^2 + (90-100)^2 + \cdots + (120-100)^2}{4} = 250 = 15.8^2$$

og

$$\text{std}(X) = 15.8$$

mens for Y

$$\text{var}(Y) = \frac{(20-100)^2 + (40-100)^2 + \cdots + (180-100)^2}{4} = 4000 = 63.25^2$$

$$\text{std}(Y) = 63.25$$

2.3 Tsjebytsjevs regel

"Sammenhengen mellom ei datamengde, gjennomsnitt og standardavik."

Minimum 75% av observasjonane har verdi i intervallet

$$\left[\overline{X} - 2s_X, \overline{X} + 2s_X\right]$$

3 Lecture 2

Vi begynner med begrepet sannsyn.

Mål:

- kjennskap til sentrale begrep
- rekne ut sannsyn
- kjenne til grunnleggande begrep frå mengdelære

3.1 Stokastisk forsøk, utfallsrom og hending

Definition 3.1: Stokastisk forsøk

- Vi veit kva utfall som er mulig
- Berre eitt utfall kan skje i kvart forsøk
- Veit ikkje utfallet vil bli

Example 3.1

Terningskast er eit eksempel på eit stokastisk forsøk. Vi veit at vi får $1-6$. Vi kan kunn få eit utfall. Vi veit ikkje på førhånd kva vi får.
Vi kaster 2 terninger. Dei mulige utfalla er

$$\begin{array}{cccccc} 1,1 & 1,2 & 1,3 & 1,4 & 1,5 & 1,6 \\ 2,1 & 2,2 & 2,3 & 2,4 & 2,5 & 2,6 \\ 3,1 & 3,2 & 3,3 & 3,4 & 3,5 & 3,6 \\ 4,1 & 4,2 & 4,3 & 4,4 & 4,5 & 4,6 \\ 5,1 & 5,2 & 5,3 & 5,4 & 5,5 & 5,6 \\ 6,1 & 6,2 & 6,3 & 6,4 & 6,5 & 6,6 \end{array}$$

som utgør $6^2 = 36$ mulige utfall.

Example 3.2

Kaster eit kronestykke 3 gongar. Dei mulige utfalla er

$$\begin{array}{cccc} KKK & KKM & KMK & KMM \\ MKK & MKM & MMK & MMM \end{array}$$

som utgjør $2^3 = 8$ mulige utfall.

Definition 3.2: Utfallsrom og hending

Utfallsrommet S er mengda av alle mulige utfall.
Hending er eitt eller fleire utfall som oppfyller ein gitt betingelse.

Example 3.3

Viss vi tar utgangspunkt i Example 3.1 kan vi se på hendinga $A = \{\text{sum er 4}\} = \{(3,1),(2,2),(1,3)\}$.
$B = \{\text{minst eit partal}\} = \{(1,2),(1,4),(1,6),\ldots\}$ som gir $|B| = 27$.

Example 3.4

Viss ein målar høgda til personer er tilstandsrommet $S = \{\text{alle reelle tal mellom 0.5 og 2.5}\}$.

3.2 Sannsyn for handling

- Bruke modell for sannsyn
- Empiri
- Magefølelse

Notasjon.

La A være ei hending:

- sannsyn for A : $P(A)$
- $0 \leq P(A) \leq 1$

Definition 3.3: Uniform sannsynlighetsmodell

La S være utfallsrommet til eit stokastisk forsøk, og la A være ei hending.
Viss alle utfall er like sannsynlige, då er

$$P(A) = \frac{\text{antal utfall der } A \text{ er gunnstig}}{\text{antal mulige utfall}}$$

Example 3.5

Betrakt terningskast eksempellet fra tidligere.
La S være terningskast med to terninger.

$$A = \{\text{sum er 4}\}$$

$$P(A) = \frac{3}{6^2} = \frac{3}{36} = \frac{1}{12}$$

$$B = \{\text{minst eitt partal}\}$$

$$P(B) = \frac{27}{36} = \frac{3}{4}$$

Empirisk sannsyn.

Kva er sannsynligheita for at ein person er mellom 175cm og 180cm?

Dette er ikkje ein uniform modell siden det ikkje er like stor sannsynligheit for ein person å ha ein gitt høgde.

Det vi kan gjere da er å velge 1000 personer, måle høgda og lage ein frekvenstabell.

Da ville vi hatt

$$P(\text{mellom 175 og 180}) = \text{ relativ frekvens}$$

Store tals lov.

Viss eit stokastisk forsøk blir gjentatt mange gangar så vil den empiriske sannsynligheita nærme seg den teoretiske sannsynligheita.

3.3 Mengdelære

Example 3.6

$A = \{\text{sum} = 4\}$, $P(A) = \frac{1}{12}$.
$B = \{\text{minst eitt partal}\}$, $P(B) = \frac{3}{4}$.
$C = \{\text{sum} \geq 9\}$, $P(C) = \frac{5}{18}$.
$P(\text{sum } = 4 \text{ og minst eitt partal})$?
$P(\text{sum})4 \text{ og sum} \geq 9)$?

Definition 3.4: Mengde

Ei mengde A er ei samling element (fra ei utvalgsmengde S).

Definition 3.5: Union, snitt, komplement, disjunkt

La A, B være mengder med element fra ei utvalgsmengde S.
Union.

$$A \cup B = \{x : x \in A \vee x \in B\}$$

Snitt.

$$A \cap B = \{x : x \in A \wedge x \in B\}$$

Komplement.

$$\overline{A} = A^c = S \setminus A = \{x \in S : x \notin A\}$$

Disjunkt.
A og B er disjunkte viss

$$A \cap B = \emptyset$$

4 Lecture 3

4.1 Grunnreglar for sannsynlighet

La S være utfallsrommet til eit stokastisk forsøk, og la A, B være hendingar.

1. $0 \leq P(A) \leq 1$
2. $P(S) = 1, P(\emptyset) = 0$
3. Viss A og B er disjunkte hendingar ($A \cap B = \emptyset$), då er $P(A \cup B) = P(A) + P(B)$
4. $P(\overline{A}) = 1 - P(A)$

Example 4.1

Betrakt kastingen av to terninger.
$A = \{\text{summen er 4}\}$
$B = \{\text{minst 1 terning er partal}\}$
$C = \{\text{sum } \geq 9\}$
$P(A) = \frac{3}{36}, P(B) = \frac{27}{36}, P(C) = \frac{10}{36}$
$P(A \cup C) = P(A) + P(C) = \frac{13}{36}$ siden A og C er disjunkte (ingen tall x oppfyller $x = 4 \wedge x \geq 9$ samtidig).
$P(A \cup B)$. Her kan vi ikke bruke regel 3, siden $2 + 2 = 4$ og $(2,2)$ oppfyller at minst en er partal, så $A \cap B = \{(2,2)\} \neq \emptyset$. I dette tilfellet har vi $P(A \cup B) = P(A \cup B) - P(A \cap B) = \frac{29}{36}$ (som addisjonsregelen for ikke-disjunkte mengder fra MAT210).
$P(\overline{A \cup B}) = 1 - P(A \cup B) = \frac{7}{36}$

4.2 Betinget sannsynlighet

Example 4.2

Eit bilfirma har 2000 tilsette.
La
$A = \{\text{stemmer Ap}\}$
$B = \{\text{er bilmekaniker}\}$
La oss si at vi vet at en person er bilmekaniker, hva er da sjansen for at de stemmer Ap?
Viss vi vet at blant bilmekanikere så stemmer 500 Ap og 380 ikke da er sjansen for at, gitt en bilmekaniker, de stemmer Ap

$$P(A|B)) \frac{500}{500 + 380} = \frac{500}{880}$$

Den betinget sannsynligheten for at A skjer viss B har skjedd

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Fra dette følger det et par formler

1. $P(A \cap B) = P(A|B) \cdot P(B)$

2. $P(A) = P(A \cap B) + P(A \cap \overline{B})$
3. $P(A) = P(A|B) \cdot P(B) + p(A|\overline{B}) \cdot P(\overline{B})$

Theorem 4.1: Bayes teorem

La A, B være hendingar da har vi at

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Kort bevis.

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\ &= \frac{\frac{P(B \cap A)}{P(A)} P(A)}{P(B)} \\ &= \frac{P(B \cap A)}{P(B)} \\ &= \frac{P(A \cap B)}{P(B)} \\ &= P(A|B) \end{aligned}$$

□

Sensitivitet, spesifisitet, basisrate.

La

B = {pasient er syk}

A = {testmetode er positiv}

Vi har som mål å finne sjansen for at en person er syk når testen er positiv, $P(B|A)$.

- Sensitivitet: $P(A|B)$, sannsynligheten for at test er positiv når pasient er syk
- Spesifisitet: $P(\overline{A}|\overline{B})$, sannsynlighet for at test er negativ når pasient er frisk
- Basisrate: $P(B)$, sannsynlighet for at gitt pasient er syk

$$\begin{aligned} P(B|A) &= P(A|B)\frac{P(B)}{P(A)} \\ P(A) &= P(A|B) \cdot P(B) + P(A|\overline{B})P(\overline{B}) \\ P(\overline{B}) &= 1 - P(B) \\ P(A|\overline{B}) &= 1 - P(\overline{A}|\overline{B}) \\ \Rightarrow P(B|A) &= P(A|B)\frac{P(B)}{P(A|B) \cdot P(B) + (1 - P(\overline{A}|\overline{B}))(1 - P(B))} \end{aligned}$$

5 Lecture 4

5.1 Diskret synnsynlighetsmodell

Mål:

- kva er ein stokastisk variabel
- innholdet i sannsynlighetsmodell
- fordelingsfunksjonar og anvendelse
- sentralmål og spredningsmål for diskret sannsynlighetsmodell

Definition 5.1: Stokastisk variabel

Ein stokastisk variabel X er ein tilfeldig variabel som knytter alle utfall i utfallsrommet til ein verdi.

Definition 5.2: Verdimengde

Dei moglege verdiane av eit stokastisk forsøk kalles verdimengda V_X.

Example 5.1

Anta at vi kaster 2 terninger.
La $X = \{\text{summen av terningene}\}$.
$X = 4 = \{13, 22, 31\}$.
$X \leq 4 = \{11, 12, 21, 13, 22, 31\}$.
$V_X = \{2, 3, 4, \ldots, 9, 10, 11, 12\}$.
$Y = \{\text{antall partall}\}$.
$V_Y = \{0, 1, 2\}$.
I eit pengespill taper du 10 viss sum ≤ 6, vinner 8 viss sum $= 7$, og vinner 1 viss sum ≥ 8.
Viss vi lar Z betegne den stokastiske variabelen i pengespillet får vi:
$V_Z = \{-10, 8, 1\}$.

Definition 5.3: Sannsynlighetsfordeling

Ein sannsynlighetsfordeling til ein stokastisk variabel X inneholder:

1. verdimengden V_X
2. sannsynligheten for alle mulige utfall i V_X

V_X	$P(X=x)$	$F(x)$	$x \cdot P(X=x)$	$x^2 \cdot P(X=x)$
2	1/36	1/36	2/36	4/36
3	2/36	3/36	6/36	18/36
4	3/36	6/36	12/36	43/36
5	4/36	10/36	20/36	100/36
6	5/36	15/36	30/36	180/36
7	6/36	21/36	42/36	294/36
8	5/36	26/36	40/36	320/36
9	4/36	30/36	36/36	324/36
10	3/36	33/36	30/36	300/36
11	2/36	35/36	22/36	242/36
12	1/36	36/36	12/36	144/36
			$E(X) = 252/36 = 7$	$E(X^2) = 1974/36$

er hvordan synnsynlighetsfordelingen kan se ut for $X = \{\text{summen av terningene}\}$. $F(x)$ er definert som sannsynligheten for at $X \leq x$. Med andre ord

$$F(x) = P(X \leq x) = \sum_{y \in V_X, y \leq x} P(X = y)$$

Example 5.2

Finn sannsynlighet for at sum er 7 eller mindre. Da skal vi finne $P(X \leq 7)$ som næmlig er

$$F(7) = \frac{21}{36}$$

Finn sannsynlighet for at sum er større enn 8.

$$P(X > 8) = 1 - P(x \leq 8) = 1 - F(8) = \frac{10}{36}$$

Sentralmål og spredningsmål.

Forventningsverdi:

$$E(X) = \mu_X = \sum_{x \in V_X} xP(X = x)$$

$$Var(X) = \sigma_X^2 = \left(\sum_{x \in V_X} x^2 P(X = x) \right) - \mu_X^2$$

$$std(X) = \sigma_X$$

6 Lecture 5

6.1 Fleire stokastiske variabler

I eit terningspel kaster ein to terningar. Det er 2 måtar å vinne.

I spel 1 vinner ein 2 viss summen av terningane er 7, vinner 1 viss sum ≥ 8 og ein vinner 0 viss sum ≤ 6.

I spel 2 vinner ein 2 viss begge terningane er partal, 1 viss ein av terningane er partal og 0 viss begge er oddetal.

La $X = \{\text{gevinst spel 1}\}$, $V_X = \{0,1,2\}$. $Y = \{\text{gevinst spel 2}\}$, $V_Y = \{0,1,2\}$.

	$y=0$	$y=1$	$y=2$	$P(X=x)$
$x=0$	1/6	1/6	1/12	5/12
$x=1$	1/12	1/6	1/6	5/12
$x=2$	0	1/6	0	1/6
$P(Y=y)$	1/4	1/2	1/4	1

$E(X), E(Y), Var(X), Var(Y)$

$$E(X) = \sum xP(X=x) = \mu_X$$

$$Var(X) = \sum x^2 P(X=x) - \mu_X^2$$

$$E(X) = 0 \cdot \frac{5}{12} + 1 \cdot \frac{5}{12} + 2 \cdot \frac{1}{6} = \frac{3}{4} = \mu_X$$

$$Var(X) = 0 + \frac{5}{12} + 4 \cdot \frac{1}{6} - \mu_X^2 = \frac{13}{12} - \mu_X^2 = \frac{13}{12} - \frac{9}{16} = \frac{25}{48} = \left(\frac{5\sqrt{3}}{12}\right)^2 = \sigma_X^2$$

$$E(Y) = \frac{1}{2} + \frac{2}{4} = 1 = \mu_Y$$

$$Var(Y) = \frac{1}{2} + \frac{4}{4} - \mu_Y^2 = \frac{1}{2} = \frac{1}{\sqrt{2}} = \sigma_Y^2$$

Interessante spørsmål.

- Viss eg spelar på begge spela, kva er forventa verdi og varians?
- Viss eg vinner på spel 1, aukar eg eller minkar eg sjansen for å vinne på spel 2?
- Er spela uavhengige av kvarandre?

La $Z = X + Y$ (den samla gevinsten ved å spele på begge spela). Verdimengda $V_Z = \{0,1,2,3,4\}$.

z	$P(Z=z)$	$zP(Z=z)$	$z^2P(Z=z)$
0	1/6	0	0
1	1/4	1/4	1/4
2	1/4	1/2	1
3	1/3	1	3
4	0	0	0
		$\mu_Z = 7/4$	$E(Z^2) = 17/4$

$E(Z) = \mu_Z = 7/4$

$Var(Z) = 17/4 - (7/4)^2 = (\sqrt{1}9/4)^2 = \sigma_Z^2$

Theorem 6.1

La X og Y vere stokastiske variablar. La $Z = aX + bY + c$.

$$E(Z) = aE(X) + bE(X) + c$$

6.2 Kovarians og korrelasjon

- Er det samanheng mellom X og Y?
- Er X og Y uavhengige?
- Vil seier i X påvirke Y?

Kovarians

$$Cov(X,Y) = E(X \cdot Y) - \mu_X \cdot \mu_Y = \left(\sum_{x,y} x \cdot yP(X = x \wedge Y = y)\right) - \mu_X\mu_Y$$

Korrelasjon

$$Corr(X,Y) = \rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}$$

Viss vi betrakter eksempellet fra tidligere får vi

$$E(X \cdot Y) = \frac{1}{6} + 2\frac{1}{6} + 2\frac{1}{6} = \frac{5}{6}$$

så

$$Cov(X,Y) = \frac{5}{6} - \mu_X\mu_Y = \frac{1}{12}$$

Deretter finner vi korrelasjon ved

$$Corr(X,Y) = \frac{1/12}{\sigma_X\sigma_Y} = \frac{1/12}{\left(\frac{5\sqrt{3}}{12}\right)\left(\frac{1}{\sqrt{2}}\right)} = \frac{\sqrt{2}}{5\sqrt{3}} = 0.164$$

For korrelasjon har vi at

$$-1 \leq Corr(X,Y) \leq 1$$

hvor nærmere -1 betyr mer negativ relasjon og nærmere 1 betyr mer positiv relasjon. Med andre ord ρ nær 1 betyr ein lineær samanheng mellom X og Y (altså en økning i X betyr en økning i Y og vice-versa). På samme måte betyr ρ nær -1 at en økning i en er en minking i den andre. $\rho = 0$ betyr ingen lineær samanheng.

Theorem 6.2

For stokastiske variabler X og Y og konstanter a, b, c har vi:

$$Var(aX + bY + c) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y)$$

Definition 6.1: Uavhengighet

To stokastiske variablar X og Y er uavhengige dersom

$$P(X = a \wedge Y = b) = P(X = a) \cdot P(Y = b)$$

Theorem 6.3

Viss to stokastiske variablar X og Y er uavhengige så er

$$Corr(X, Y) = 0$$

Denne implikasjonen går bare en veg.

7 Lecture 6

7.1 Binomiske forsøk

Example 7.1

Ei skål inneholder 20 kuler. 8 av kulene er raude og 12 er ikkje raude. Vi trekker 5 kuler ut av skåla, og legger tilbake kula. Kva er sannsynligheita for å trekke 3 raude kuler?
La R - kula raud, $P(R) = \frac{2}{5}$.
$\overline{R}$ - kula ikkje raud, $P(\overline{R}) = \frac{3}{5}$.
Sannsynligheten for å trekke 3 raude er då sannsynligheita for å trekke raud 3 gongar og sannsynligheita for å trekke ikkje raud 2 gongar, siden hendelsane ikkje er relaterte (tilbakelegging).
Dermed er

$$P(RRR\overline{R}\overline{R}) = \left(\frac{2}{5}\right)^3 \left(\frac{3}{5}\right)^2 = \frac{72}{3125}$$

men dette er bare en av måtene å trekke 3 raude.
Vi kunne også tatt

$$P(RR\overline{R}R\overline{R}) = \frac{72}{3125}$$

Som vi ser, sannsynligheten for å trekke 3 raude er uavhengig av rekkefølgen. Dermed kan vi legge sammen alle sannsynligheitene for alle måtene å velge 3 raude og 2 ikkje raude.
Måtene å velge 3 raude er

$$\binom{5}{3} = 20$$

Dermed er sjansen for å trekke 3 raude er

$$20 \cdot \frac{72}{3125}$$

Definition 7.1: Binomisk forsøk

1. Det utføres n forsøk
2. Kvart forsøk har to utfall A eller $\overline{A}$
3. Sannsynet $P(A)$ er likt i kvart forsøk
4. Forsøka er uavhengige

Viss X - antal ganger A skjer er X ~binomial (n, p).

$$P(X = x) = \binom{n}{x} P^x (1-p)^{n-x}$$

$$E(X) = np$$

$$Var(X) = np(1-p)$$

Example 7.2

Trekker kuler ut av ei skål. Legger kula tilbake. 8 raude, 12 ikkje raude. Trekker 5 kuler.
X ~antall raude kuler, så er det binomial(5, $\frac{2}{5}$).
$E(X) = 5 \cdot \frac{2}{5} = 2$
$Var(X) = 5 \cdot \frac{2}{5}(1 - \frac{2}{5}) = 1.2$.

$$p(X = 0) = \binom{5}{0}\left(\frac{2}{5}\right)^0\left(\frac{3}{5}\right)^5 = 0.074$$

$$P(X = 1) = \binom{5}{1}\left(\frac{2}{5}\right)^1\left(\frac{3}{5}\right)^4 = 0.259$$

$$P(X = 2) = \binom{5}{2}\left(\frac{2}{5}\right)^2\left(\frac{3}{5}\right)^3 = 0.346$$

Problem 7.1: Oppgave 5.2

Du er sulten, og kjøper åtte tilfeldig valgte pølser. Sannsynet for at pølsa er sprukken er 0.2 for hver av pølsene. La X være antall sprukne pølser blant de 8. Hvilken sannsynlighetsfordeling for X? Finn $P(X = 2)$.

Løsning. X er et binomisk forsøk siden hver utfallet er sprukken eller ikke sprukken. Det at en pølse er sprukken påvirker ikke sannsynet for at en annen pølse er sprukken, så forsøka er uavhengige. Dermed er X ~binomial(8,0.2).

$$P(X = 2) = \binom{8}{2}(0.2)^2(0.8)^6 = 0.29360128$$

litt mindre en 1/3 gongar. □

Med utgangspunkt i Oppgave 5.2, hva er da sjansen for at mellom 2 og 5 av pølsene sprekker?

$$\begin{aligned} P(2 \le X \le 5) &= \sum_{x=2}^{5} P(X = x) \\ &= \binom{8}{2}0.2^2 0.8^6 + \cdots + \binom{8}{5}0.2^5 0.8^3 \end{aligned}$$

som blir en ganske lang utregning. Kan vi være mer effektive?

Fra *Kapittel 4.1* har vi

$$P(a \le X \le b) = P(X \le b) - P(X < a)$$

og

$$P(X \ge a) = 1 - P(X < a)$$

Så

$$P(2 \le X \le 5) = P(X \le 5) - P(X \le 1) = 0.999 - 0.503 = 0.496$$

7.2 Poissonfordeling

Definition 7.2: Poissonprosess

La A = {antall førekomster}. Da er A ein poissonprosess viss

1. Antall gongar A intreffer i eit tidsinterval er uavhengig av andre førekomster.
2. Forventa antal førekomster i ei gitt tidsenhet er konstant.
3. To førekomster kan ikkje skje samtidig.

Theorem 7.1

X - antal gongar A skjer.
X ~poisson(λt) der λ er en konsentrasjon og t er eit tidsinterval.
Da er

$$E(X) = Var(X) = \lambda t$$

$$P(X = x) = \frac{(\lambda t)^x}{x!} e^{-\lambda t}$$

Example 7.3

Ved en bensinstasjon er det i snitt 10 bilar som fyller bensin per time. Da er $\lambda = \frac{10 \text{ bilar}}{t}$.
La Y = {antall tankingar på 30 min}. Vi antar at raten ikke endrer seg, at forventet antal biler per time er konstant og at to biler ikke kan tanke på same tid.
Da er Y en poissonprosess med $t = 0.5$, så Y ~poisson(5).

$$P(Y = 10) = \frac{5^{10}}{10!} e^{-5} = 0.018$$

$$P(Y = 4) = \frac{5^4}{4!} e^{-5} = 0.175$$

$$P(Y \leq 8) = \sum_{y=0}^{8} P(Y = y)$$

Dette er langt å rekne ut så vi bruker en tabell for å finne

$$P(Y \leq 8) = 0.932$$

Et viktig punkt er at λ og t ikkje treng å være tidsavhengig. For eksempel kan λ være antall mol/liter og t være volum. Eller λ sauar/m^2 og t være m^2, osv.

Example 7.4

I prøvetaking av avløpsvatn kan du spore molekyla i vatnet. Vi forventer 100 molekyl fentanyl pr.liter. Ein tank inneholder $40mL$.
X = antall molekyl = poisson$(100 \cdot 0.04)$ = possion(4)
$E(X) = 4$ molekyl med fentanyl.
$Var(X) = 4$, så 90% av tiden har vi mellom 2 og 6 molekyl med fentanyl.

$$P(X = 8) = \frac{(\lambda t)^x}{x!} e^{-\lambda t} = \frac{4^8}{8!} e^{-4} = 0.030$$

så det er 3% sjanse å finne 8 molekyl fentanyl i en $40mL$ prøve.

$$\begin{aligned} P(3 \leq X \leq 5) &= P(X \leq 5) - P(X \leq 2) \\ &= 0.547 \end{aligned}$$

via poisson-tabell.

8 Lecture 7 | Kontinurelige sannsynlighetsmodellar

Definition 8.1: Sannsynlighetstetthet

Sannsynlighetstettheten $f(x)$ beskriver fordelinga av ein kontinuerlig variabel, og oppfyller:

1. Arealet under kurven er 1.
$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$
2. Sannsynligheten for utfall mellom a og b er gitt ved
$$P(a \leq X \leq B) = \int_{a}^{b} f(x)\,dx$$
3. $f(x) \geq 0$

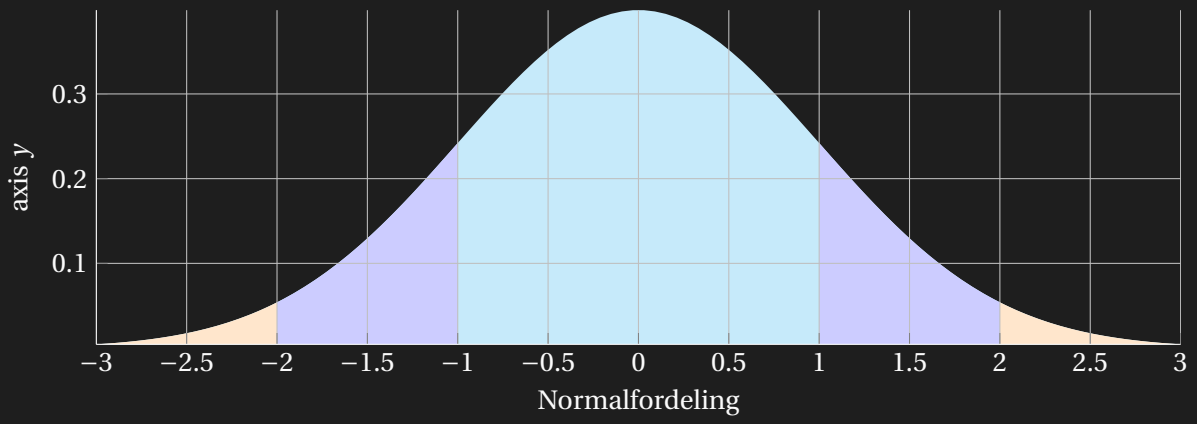
Definition 8.2: Fordelingsfunksjon

La
$$F(x) = \int_{-\infty}^{x} f(x)\,dx$$
Da er $F(a)$ sannsynligheten for at X er mindre enn eller like a, i.e. $P(X \leq a)$.
Da er $1 - F(a)$ sannsynligheten for at X er større enn a, i.e. $P(a < X)$.
$F(B) - F(A)$ er sannsynligheten for $P(a < X \leq b)$.

Definition 8.3: Normalfordelt

Ein variabel X er normalfordelt dersom med middelverdi $E(X) = \mu$ og standardavik σ.
$$X \sim \mathcal{N}(\mu, \sigma)$$
viss tetthetsfunksjonen er
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Tetthetsfunksjonen beskrivd i 8.3 ser slik ut:



Her har vi $\mu = 0$ som fører til at høydepunktet på kurven er ved 0.

Regel.

Anta at $X \sim \mathcal{N}(\mu, \sigma)$.

Da er

$$P(X \leq b) = G\left(\frac{b-\mu}{\sigma}\right)$$

$$P(a \leq X \leq b) = G\left(\frac{b-\mu}{\sigma}\right) - G\left(\frac{a-\mu}{\sigma}\right)$$

og

$$P(X > b) = 1 - G\left(\frac{b-\mu}{\sigma}\right)$$

Hvor vi leter opp

$$G\left(\frac{b-\mu}{\sigma}\right)$$

i en tabell.

Example 8.1

Anta at høgda av 18 åringar er $\mathcal{N}(180,7)$.

- Kva er sannsynlighet for å velge en person mellom 175- og 185cm?

$$\begin{aligned} P(175 \leq X \leq 185) &= G\left(\frac{5}{7}\right) - G\left(-\frac{5}{7}\right) \\ &= G(0.71) - G(-0.71) \\ &= 0.7611 - 0.2389 \\ &= 0.52 \end{aligned}$$

- Kva er sannsynligheten for å velge ein person høgare enn 200cm?

$$\begin{aligned} P(X \geq 200) &= 1 - G\left(\frac{20}{7}\right) \\ &= 1 - G(2.86) \\ &= 1 - 0.9979 \\ &= 0.0021 \end{aligned}$$

- Finn eit intervall $[180-x, 180+x]$ slik at det er 90% sjanse for at ein gitt person har høgde innafor intervallet.

$$\begin{aligned} P(180 - x \leq X \leq 180 + x) &= 0.9 \\ G\left(\frac{x}{7}\right) - G\left(-\frac{x}{7}\right) &= 0.9 \\ x/7 &= 1.65 \\ x &= 11.55 \end{aligned}$$

Så intervallet blir $[168.45, 191.55]$.

Theorem 8.1: Spredningsinterval

La $X \sim \mathcal{N}(\mu, \sigma)$.

- Da er det $100(1-\alpha)\%$ sikkert at utfallet av X er mindre enn $\mu + Z_\alpha \sigma$.
- Det er $100(1-\alpha)\%$ sjanse for at utfallet av X er større enn $\mu - Z_\alpha \sigma$.
- Det er $100(1-\alpha)\%$ sikkert at utfallet av X er mellom $\mu - Z_{\alpha/2}\sigma$ og $\mu + Z_{\alpha/2}\sigma$.

Regel. La $X_1, X_2, \ldots, X_n$ være uavhengige. Anta $X_i \sim \mathcal{N}(\mu, \sigma)$. Da er summen av forsøka $X_1 + X_2 + \cdots + X_n \sim \mathcal{N}(n\mu, \sqrt{n}\sigma$. Da er gjennomsnittet $\frac{1}{n}(X_1 + \cdots X_n) \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

Example 8.2

Anta øretvekt er $\mathcal{N}(300, 50)$.
En dag fekk eg 50 fisk. Kva er sannsynligheten for at totalvekta er større en 15750g?
Y = totalvekt = $\sum_{i=1}^{50} X_i \sim \mathcal{N}(50 \cdot 300, \sqrt{50} \cdot 50) = \mathcal{N}(15000, 354)$.

$$\begin{aligned} P(Y > 15750) &= 1 - G\left(\frac{15750 - 15000}{354}\right) \\ &= 1 - G(2.12) \\ &= 1 - 0.983 \\ &= 0.017 \end{aligned}$$

Theorem 8.2: Sentralgrenseteoremet

La $X_1, \ldots, X_n$ være uavhengige forsøk frå samme sannsynlighets-fordeling. Anta $E(X_i) = \mu$, $Var(X_i) = \sigma^2$. Viss n er tilstrekkelig stor (tommelregel $n \geq 30$),

$$\frac{\sum_{i=1}^{n} X_i}{n} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

og

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, \sqrt{n}\sigma)$$

Example 8.3: Galton

La X_i = antall høgrehopp ~binomial(30, 0.5).
$E(X) = n \cdot p = 15$.
$Var(X) = np(1-p) = 2.74^2$.
Anta at vi slipper 1000 kuler. La $Y = \sum_{i=1}^{1000} X_i \sim \mathcal{N}(15000, 86.6)$.

9 Lecture 8 | Konfidensinterval

Example 9.1

La oss si at vi har to fjell hvor vi vil måle avstanden mellom toppene av desse fjellene.
Da vil det være mange faktorer som påvirker nøyaktigheten av målingen som f.eks feil i måleutstyr, feil av den som måler avstanden, og mer.
Vi øsnker å kunne finne et intervalt slik at vi kan være x antall % sikker på at høyden faller innenfor det intervalet. I vårt eksempell ønsker vi å finne et interval slik at det er 95% sikkerheit for at den faktiske høgda ligger i intervalet.

- Eg veit usikkerheita til måleutstyret av målinga $X \sim \mathcal{N}(\mu, 30)$, hvor μ (middelverdien) forventer vi at er den nøytaktige verdien og 30 (standardaviket) er usikkerheita.

Regel 5.16 (Repitisjon).
Viss $X \sim \mathcal{N}(\mu, \sigma)$ veit vi at det er $100(1-\alpha)\%$ sikker på at X er i intervallet $\mu \pm Z_{\alpha/2} \cdot \sigma$.

- $\sigma = 30$
- $100(1-\alpha) = 95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025$

  $Z_{0.025} = 1.960$

Vi trenger da eit estimat for μ. Vi måler avstanden, og det beste estimatet er resultatet av målinga.

$$X_1 = 3245m$$

Eit 95%-Konfidensinterval er gitt ved

$$\widehat{\mu} = \pm Z_{\alpha/2} \cdot \sigma$$

$$3245 \pm 1.960 \cdot 30 = [3186, 3303]$$

når vi runder av til næreste meter.
For å forbedre estimatet tar vi fleire målingar.
Anta totalt 9 målingar, der gjennomsnittet ble

$$Z = \frac{X_1 + \cdots + X_9}{9} = 3204m$$

Sentralgrenseteoremet seier då at

$$Z \sim \mathcal{N}(\mu, 30/\sqrt{9}) = \mathcal{N}(\mu, 10)$$

så usikkerheten har gått ned ved fleire målinger.
Eit 95%-konfidensinterval blir da

$$\widehat{\mu} \pm Z_{\alpha/2} \cdot (\sigma/\sqrt{n})$$

$$3204 \pm 1.96 \cdot 10 = [3184, 3223]$$

Definition 9.1: Estimator

Ein estimator $\widehat{\theta}$ er ein stokastisk variabel knytta til ei forsøksrekke $\theta_1, \theta_2, \ldots$ som estimerer θ.

1. $E(\widehat{\theta}) = \theta$
2. $Var(\widehat{\theta})$ skal være minst mulig.
3. $Var(\widehat{\theta})$ skal gå mot 0 viss antal forsøk auker.

Estimator for μ.

$$\widehat{\mu} = \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Estimator for σ.

$$\widehat{\sigma^2} = S^2 = \frac{1}{n-1}\sum (X_i - \overline{X})^2$$

Estimator for P.

$$\widehat{P} = \frac{X}{n}$$

der X er positive utfall og n er antall forsøk.

Problem 9.1

θ er ein ukjent stokastisk variabel.
Mål: Finne eit interval $[A, B]$ slik at

$$P(A \leq \theta \leq B) = 1 - \alpha$$

Z-interval.

- μ - ukjent
- σ - kjent

Da er eit $100(1-\alpha)\%$ konfidensinterval gitt ved

$$X \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

P-interval.

Da er eit $100(1-\alpha)\%$ konfidensinterval gitt ved

$$\widehat{P} \pm Z_{\alpha/2}\sqrt{\frac{\widehat{P}(1-\widehat{P})}{n}}$$

T-interval.

- μ - ukjent
- σ - ukjent

Eit $100(1-\alpha)\%$-konfidensinterval for μ er gitt ved

$$\overline{X} \pm t_{\alpha/2} \cdot \frac{S}{\sqrt{n}}$$

der S er estimatet for standardaviket, og $t_{\alpha/2}$ har $(n-1)$ frihetsgrader.

- Siden σ er ukjent er eit T-interval større enn eit Z-interval.
- Viss $n > 30$ er T-interval tilnærma lik Z-interval.

Problem 9.2: 5.13

X = diametere $\sim \mathcal{N}(\mu, \sigma)$, $n = 6$.
Vi ønsker 95%-konfidensinterval $\Rightarrow \alpha/2 = 0.025$.

$t_{0.025}$ med 5 frihetsgrad = 2.571.

$\overline{X} = (31+32+30+31+29+30)/6 = 30.5$

$S^2 = \frac{1}{5}((31-30.5)^2 + (32-30.5)^2 + \cdots + (30-30.5)^2) = 1.1 \approx 1.05^2$

Da blir konfidensintervalet

$$30.5 \pm 2.571 \cdot \frac{1.05}{\sqrt{6}} = [29.4, 31.6]$$

Konfidens for P.

Eit $100(1-\alpha)\%$-konfidensinterval for P er gitt ved

$$\widehat{P} \pm Z_{\alpha/2} \sqrt{\frac{\widehat{P}(1-\widehat{P})}{n}}$$

Problem 9.3: 6.16

$n = 36$
$X = 29$
Da er $\widehat{P} = \frac{29}{36}$.
Vi vil finne 90%-konfidensinterval.

$$\alpha/2 = 0.05$$
$$Z_{0.05} = 1.645$$

Da blir konfidensintervalet

$$\frac{29}{36} \pm \sqrt{\frac{29/36(1-29/36)}{36}} = [0.70, 0.91]$$

10 Lecture 9 | Hypotesetesting

Example 10.1

Politiet utfører ei hastighetsmåling i ei 80 sone. Dei bruker lasermåler $X \sim \mathcal{N}(\mu, 3)$. Dei utfører 5 uavhengige målinger.

$$x_1 = 79.9$$
$$x_2 = 81.5$$
$$x_3 = 83.2$$
$$x_4 = 83.1$$
$$x_5 = 82.0$$

Tar ein gjennomsnittet av desse målingane får vi

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{3}{\sqrt{5}}\right)$$

Spørsmål.

$$\text{Kjører personen for fort?} \begin{cases} \text{ja, og han skal dommast} \\ \text{nei han blir frikjent} \end{cases}$$

Vi må ha rom for tvil:
Enten: bevise med tilstrekkelig sikkerhet at dei har kjørt for fort.
Eller: frikjennast.
Tilstrekkelig = signifikansnivå. Viss vi krever signifikansnivå $\alpha = 0.05$ betyr det i praksis at viss vi har 100 folk som kjøyrer nøyaktig fartsgrensa skal 5 personar dømmast urettferdig.
Hypoteser.
Nullhypotese H_0 (må ikkje bevisast).

$$H_0 : \mu \leq 80$$

Alternativ hypotese H_1 må bevisast.

$$H_1 : \mu > 80$$

Definition 10.1

Testobservatør.
Ein stokastisk variabel som vi baserer beslutning på.

$$\widehat{\mu} = \overline{X} = \frac{X_1 + \cdots + X_n}{n}$$
$$\widehat{P} = \frac{X}{n}$$

From av forkastningsområde.
Høgresidig: Forkast H_0 viss observatør er for stor.
Venstresidig: Forkast H_0 viss observatør er for liten.
Tosidig: Forkast H_0 viss observatlør enten er for stor eller viss den er for liten.
Signifikansnivå.
Vi må fastsette kor stor sannsynlighet vi kan tillate for å feilaktig forkaste H_0.
Styrkefunksjon.
Styrkefunksjon $\gamma(x)$ angir sannsynet for å forkaste H_0.
p-verdi.
p-verdien av ein test er sannsynet P for at ein foraster H_0 (og påstår H_1) viss H_0 er sann.

Hypotesetest(Z-test, μ ukjent, σ kjent).

La

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

Høgresidig test:

$$\begin{cases} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$

Forkast H_0 viss $Z > Z_\alpha$.

Styrkefunksjon: $\gamma(x) = 1 - G\left(Z_\alpha - \frac{X-\mu_0}{\sigma/\sqrt{n}}\right)$

Venstresidig test:

$$\begin{cases} H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$$

Forkast H_0 viss $Z < -Z_\alpha$.

Dobbeltsidig test:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

Forkast H_0 viss $|Z| > Z_{\alpha/2}$.

Hypotesetest(T-test, μ ukjent, σukjent). La

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

t_α has $n-1$ frihetsgrader.

Høgresidig:

$$\begin{cases} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$

Forkast H_0 dersom $T > t_\alpha$.

Dette fortsetter helt likt som med Z-test bare med T isteden.

Example 10.2: 6.21

Hypotese
$H_0 : \mu \leq 100$
$H_1 : \mu > 100$
Vi velger å bruke signifikansnivå $\alpha = 0.05$.
Testobservatør: $\overline{X} = \frac{103+\cdots+103}{9} = 103.8$.
Estimerer $S^2 = \frac{1}{9-1}\left((103-103.8)^2 + \cdots + (103-103.8)^2\right) = 8.695 = 2.95^2$.
$t_0.05$ med 8 frihetsgrader = 1.86.

$$T = \frac{103.8 - 100}{2.95/\sqrt{9}} = 3.86$$
$$\Rightarrow T > t_\alpha$$

Dermed forkaster vi H_0.

Hypotesetest for sannsyn P. La

$$Z = \frac{\widehat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{X - np_0}{\sqrt{np_0(1-p_0)}}$$

Høgresidig:

$$\begin{cases} H_0 : p \leq p_0 \\ H_1 : p > p_0 \end{cases}$$

Forkaster H_0 dersom $Z > z_\alpha$.

Venstresidig:

$$\begin{cases} H_0 : p \geq p_0 \\ H_1 : p < p_0 \end{cases}$$

Forkaster H_0 dersom $Z < -z_\alpha$.

Dobbeltsidig:

$$\begin{cases} H_0 : p = p_0 \\ H_1 : p \neq p_0 \end{cases}$$

Forkaster H_0 dersom $|Z| > z_{\alpha/2}$.

Example 10.3: 6.29

$n = 134$
$X = 6$
Estimator/testobservatør: $\widehat{P} = \frac{6}{134} = 0.045$
Hypotese:

$$\begin{cases} H_0 : p = 0.1 \\ H_1 : p \neq 0.1 \end{cases}$$

Signifikansnivå $\alpha = 0.05$, $z_{0.025} = 1.960$
Finner

$$Z = \frac{6 - 134 \cdot 0.1}{\sqrt{134 \cdot 0.1 \cdot 0.9}} = -2.131$$

$$|Z| = 2.131 > z_{0.025} = 1.960$$

Dermed forkaster vi H_0.

Konfidensinterval og hypotesetest.

Gitt hypotesa

$$\begin{cases} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$

Test med signifikansnivå α.

Da er $(-\infty, \mu]$ H_0 sitt gyldighetsområde og (μ, ∞) H_1 sitt gyldighetsområde.

Anta at vi har eit $100(1-\alpha)$ konfidensinterval $[x_0, x_1]$.

Dersom

$$[x_0, x_1] \subseteq (-\infty, \mu] \text{ og } [x_0, x_1] \cap (\mu, \infty) = \emptyset$$

forkaster vi ikke H_0.

Dersom

$$[x_0, x_1] \subseteq (\mu, \infty) \text{ og } [x_0, x_1] \cap (-\infty, \mu] = \emptyset$$

forkaster vi H_0.

Ellers viss

$$[x_0, x_1] \subseteq (-\infty, \mu] \text{ og } [x_0, x_1] \subseteq (\mu, \infty)$$

beholder vi fortsatt H_0.

11 Lecture 10 | Regresjon

Er det ein samanhenge mellom to stokastiske variabler X og Y?

$(X_1, Y_1), \ldots, (X_n, Y_n)$.

Spørsmål:

Er det ein lineære samanhgen mellom X og Y?

Spredningsplott. Ved å plotte datapunkt i planet, viss det er en sterk lineære samanheng mellom x- og y-verdiene finne ligninga for ei linjer som korrelerer x- og y-verdiene.

Definition 11.1: Empirisk korrelasjon

Gitt n observasjonar $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
La

$$S_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2} \text{ standardavvike for x-ane}$$

$$S_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \text{ standardavvik for y-ane}$$

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) \text{ kovarians}$$

Da er korrelasjonen

$$r = \frac{S_{XY}}{S_X S_Y}$$

Example 11.1

x	4	2	3	5	2
y	3	2	2	4	1

$\overline{X} = 3.2$
$\overline{Y} = 2.4$
$S_X^2 = 1.31^2 \Rightarrow S_X = 1.31$
$S_Y^2 = 1.14^2 \Rightarrow S_Y = 1.14$
$S_{XY} = 1.4$

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{1.4}{1.31 \cdot 1.14} = 0.94$$

Egenskaper av korrelasjon.

- $|r| \leq 1$
- Jo nærmare r er 1, jo meir indikerer det at observasjonane ligger langs ei aukande linje.
- Jo nærmare r er -1, jo meir indikerer det at observasjonane ligger langs ei minkande linje.
- Jo nærmare r er 0, jo meir indikerer det ingen lineær samanheng.

11.1 Lineær regresjon

Ved ein høg korrelasjon vil vi finne ei linje som best beskriver samanhengen mellom x og y.

Minste kvadratets rette linje.

$\widehat{y} = \widehat{\alpha} + \widehat{\beta}x$

$$\widehat{\beta} = r\frac{S_Y}{S_X}$$

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{X}$$

Med utgangspunkt i eksempelet over får vi

$$\widehat{\beta} = 0.94 \cdot \frac{1.14}{1.31} = 0.82$$

$$\widehat{\alpha} = 2.4 - 0.82 \cdot 3.2 = -0.224$$

Da får vi

$$\widehat{y} = -0.224 + 0.82x$$

Modellens godhet.

Del av variasjon blant observasjonane som blir forklart av modell r^2.

Del av variasjon som skyldes tilfeldige avvik er $1 - r^2$.