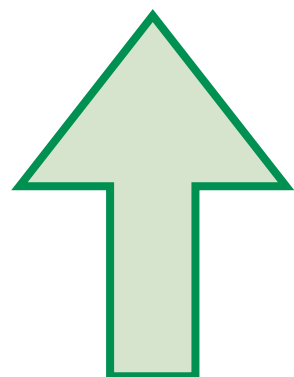
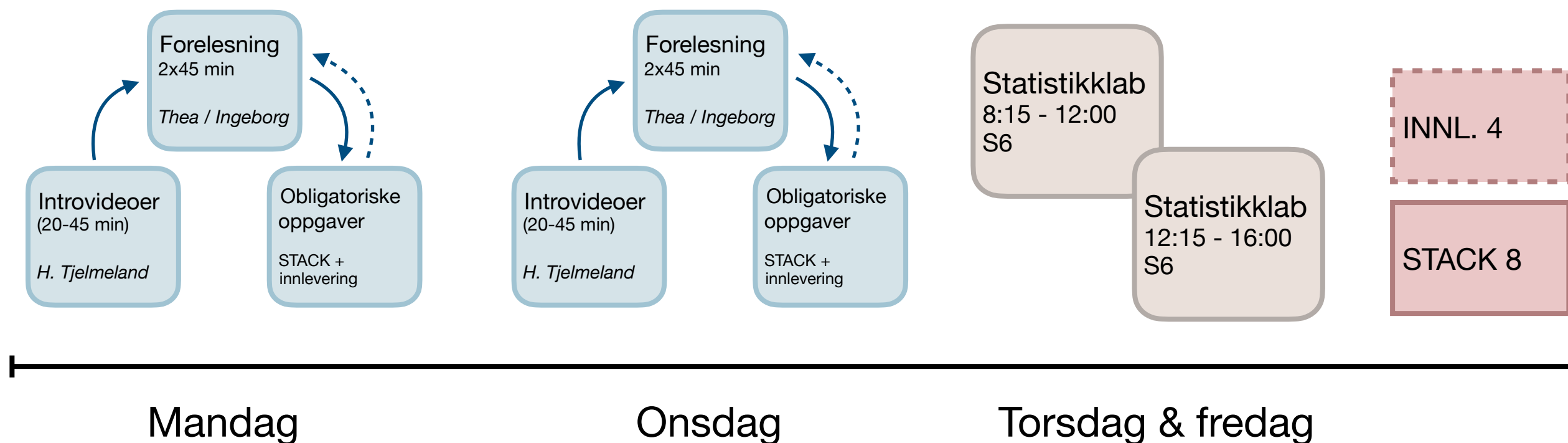


Blokk 2

TMA4245 Statistikk

Uke 9 - mandag



Uke 9 (mandag 26. februar - fredag 1. mars)



Introvideoer

Introduksjon til parameterestimering

TMA4240/TMA4245 Statistikk

Håkon Tjelmeland

Institutt for matematiske fag

Norges teknisk-naturvitenskapelige universitet

Oppgaver

STACK 8: Oppg. 1-2
INN. 4: Oppg. 1, 2

Histogram, boksplott, normalfordelingsplott og spredningsplott

TMA4240/TMA4245 Statistikk

Håkon Tjelmeland

Institutt for matematiske fag

Norges teknisk-naturvitenskapelige universitet

Observator, estimator og evaluering av estimatorer

TMA4240/TMA4245 Statistikk

Håkon Tjelmeland

Institutt for matematiske fag

Norges teknisk-naturvitenskapelige universitet

Fruktbarhetstall

“Samlet fruktbarhetstall er et mål for hvor mange barn kvinner i en befolkning i snitt vil føde dersom det gjeldende fruktbarhetsmønsteret varer ved.”
(https://snl.no/samlet_fruktbarhetstall)

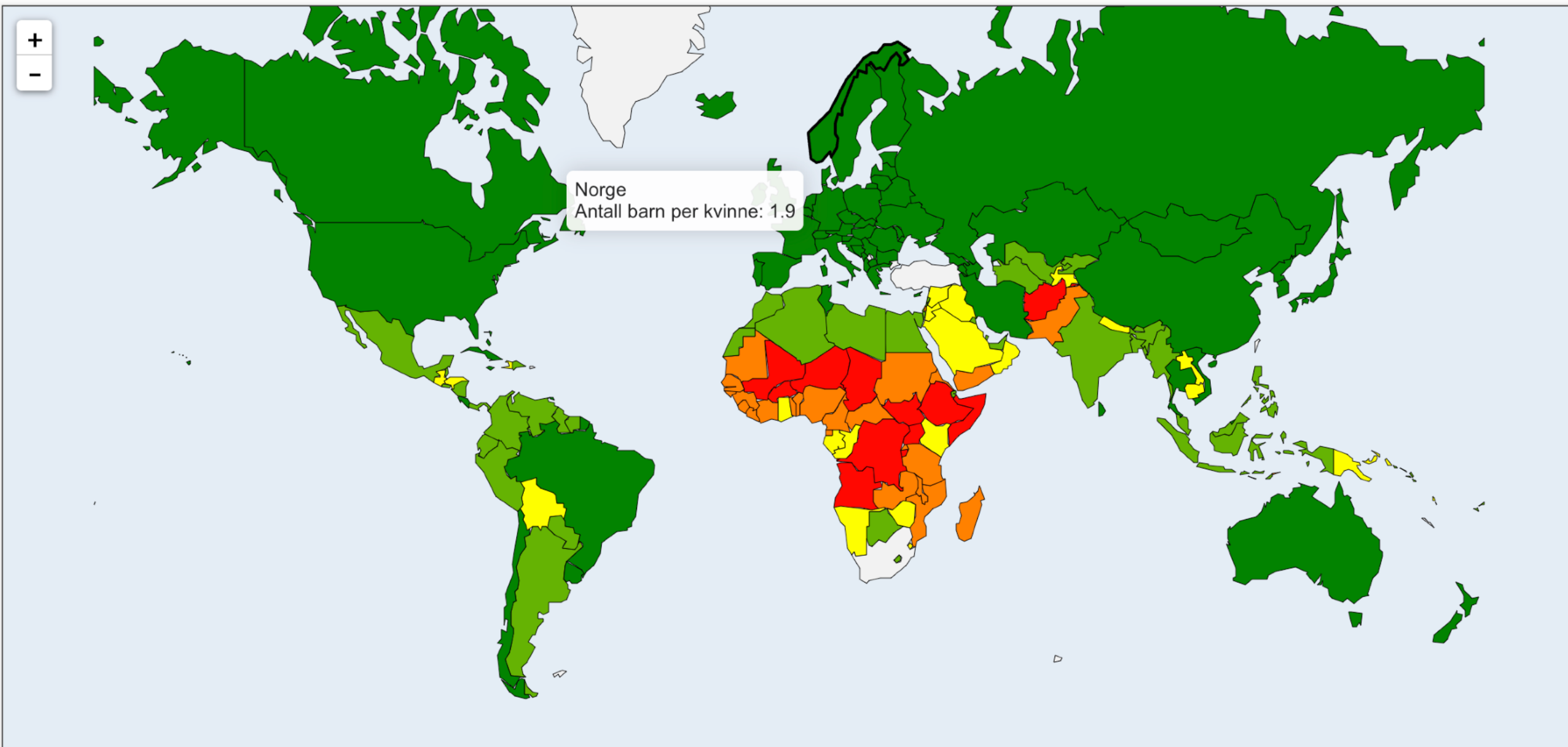
For å finne fruktbarhetstallet kan vi spørre alle norske kvinner (= hele populasjonen) om hvor mange barn de har (tidkrevende!).

Eller, vi kan estimere fruktbarhetstallet ved å spørre et tilfeldig utvalg fra populasjonen.

Gå til www.menti.com

Fruktbarhet

2000



Antall barn per kvinne



Fruktbarhetstallene viser hvor mange barn som fødes per kvinne.

Kilde: <https://www.fn.no/Statistikk/fruktbarhet>

Definisjon (Populasjon og utvalg)

En populasjon består av alle mulige observasjoner man kan gjøre. Et utvalg er en delmengde av en populasjon.

- ★ Ønsker å bruke hva vi har observert på utvalget til å få informasjon/si noe om hele populasjonen
- ★ Et utvalg må være representativt
- ★ Dersom elementene i en populasjon har fordeling $f(x)$ sier vi at vi har en $f(x)$ -populasjon

Fruktbarhet i Norge i år 2000: **1.9.**

Fruktbarhet i Norge i år 2005: **1.8.**

Fruktbarhet i Norge i år 2024: **1.5.**

Hvorfor ble vårt tall feil? (diskuter i 1-2 min)

På grunn av utvalget! Vi har ikke samlet inn data om kvinnene som ikke har barn.

Utvalget er ikke representativt for populasjonen.

Blokk 1: Sannsynlighetsregning

Kjenner $f(x; \theta)$ og verdien til θ

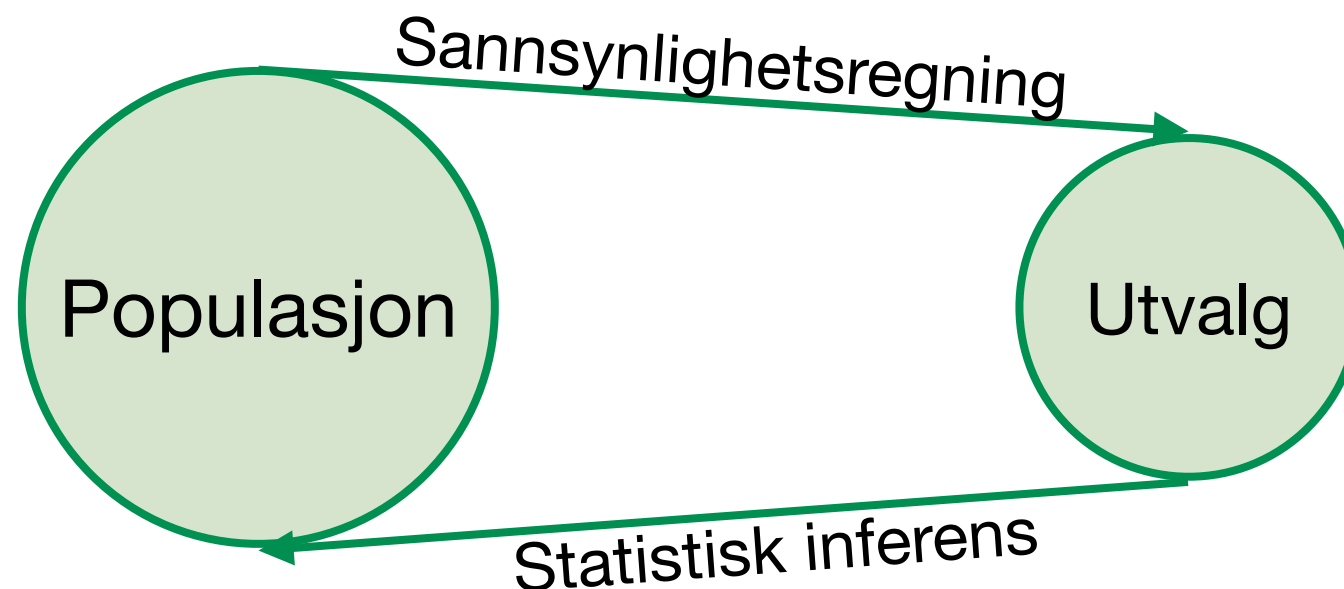
Regner på sannsynlighet for ulike hendelser

Blokk 2: Statistisk inferens

Kjenner (eller kan anta) $f(x; \theta)$,
men kjenner ikke verdien til θ

Ønsker å si noe om populasjonen
basert på et tilfeldig utvalg

Det kan vi gjøre ved bruk av
sannsynlighetsregning!



Definisjon (Tilfeldig utvalg)

La $X_1, X_2, \dots, X_n$ være n uavhengige stokastiske variabler, alle med samme fordeling $f(x)$. Vi sier da at vi har et tilfeldig utvalg fra $f(x)$ -populasjonen.

- ★ Merk: Hvis vi har et tilfeldig utvalg fra $f(x)$ -populasjonen blir simultanfordelingen til $X_1, X_2, \dots, X_n$ lik

$$f(x_1, x_2, \dots, x_n) = f(x_1) \cdot f(x_2) \cdot \dots \cdot f(x_n) = \prod_{i=1}^n f(x_i)$$

Kan “se på” utvalget:

Gjennomsnitt

Empirisk varians/standardavvik

Median

Kvartiler

Histogram

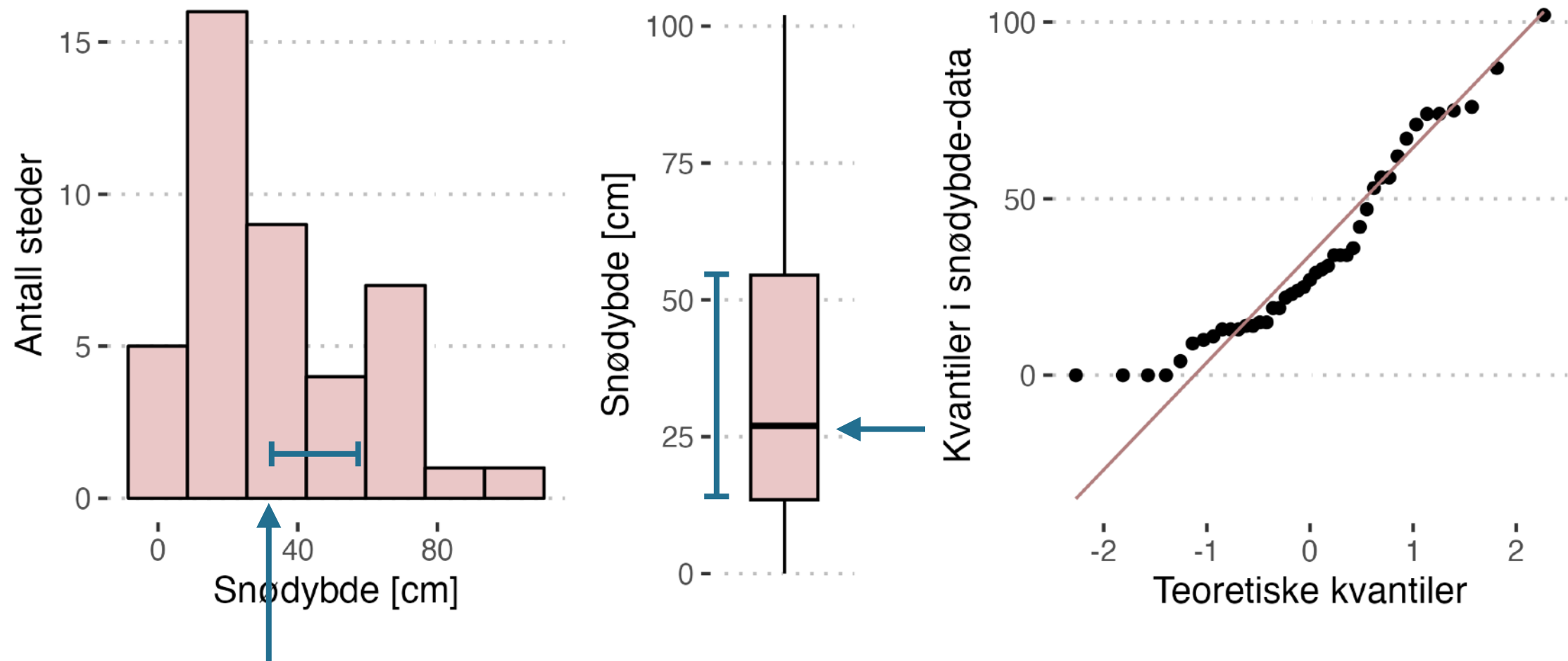
Boksplott

Normalfordelingsplott

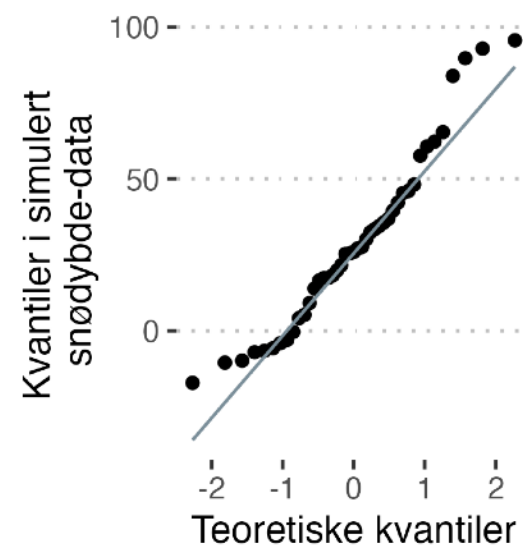
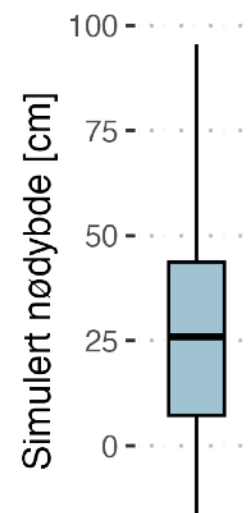
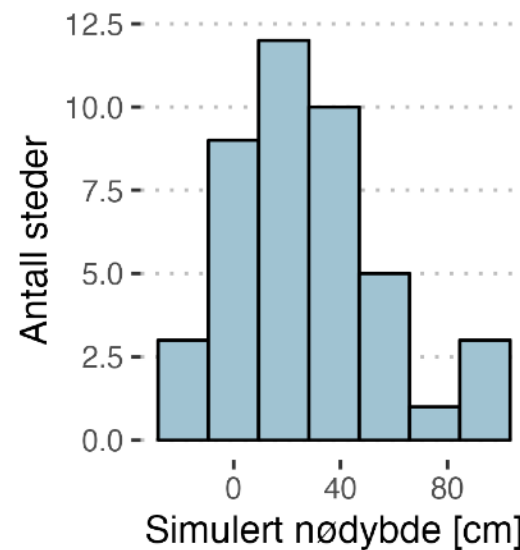
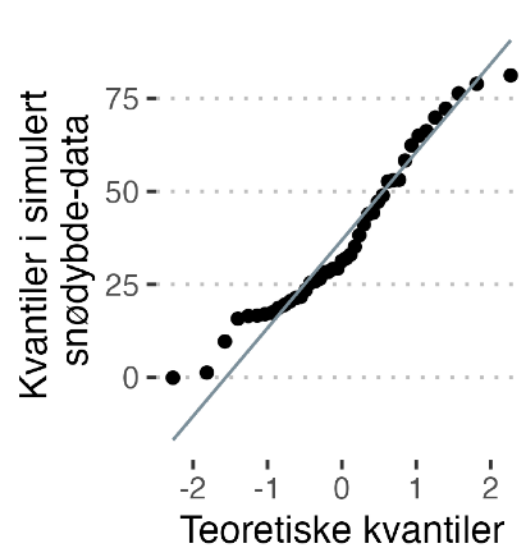
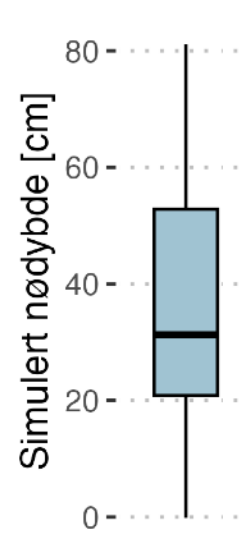
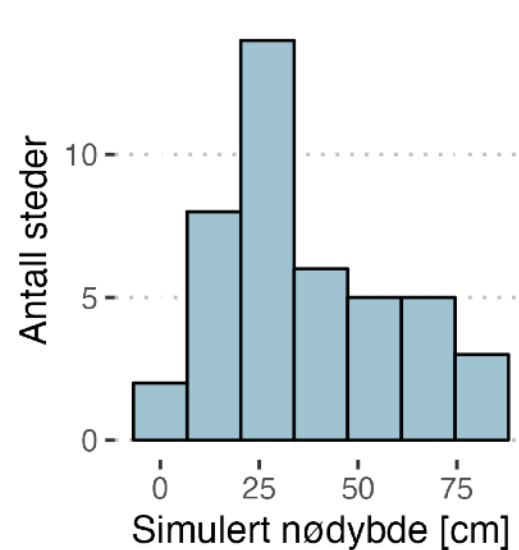
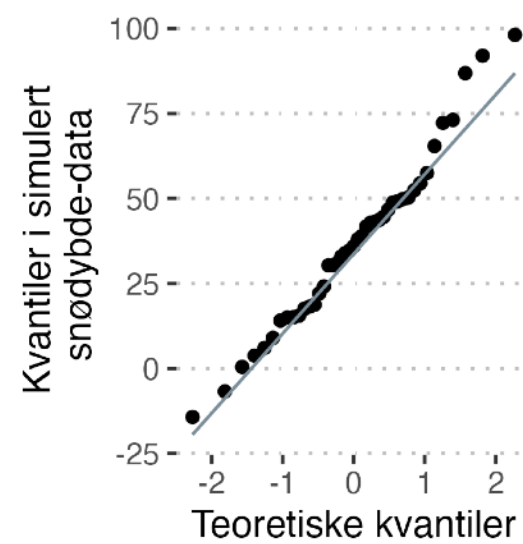
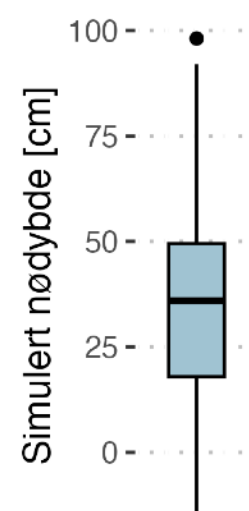
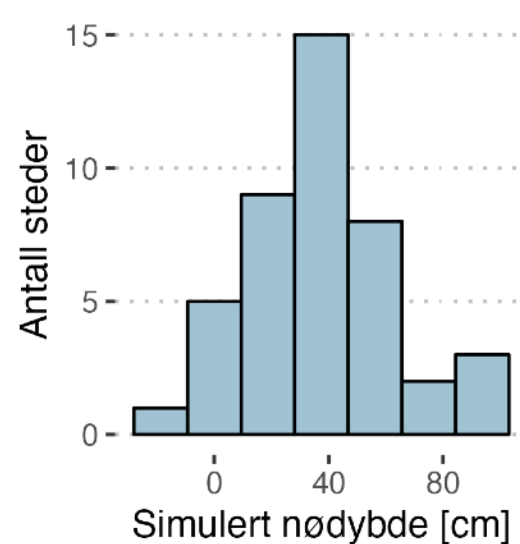
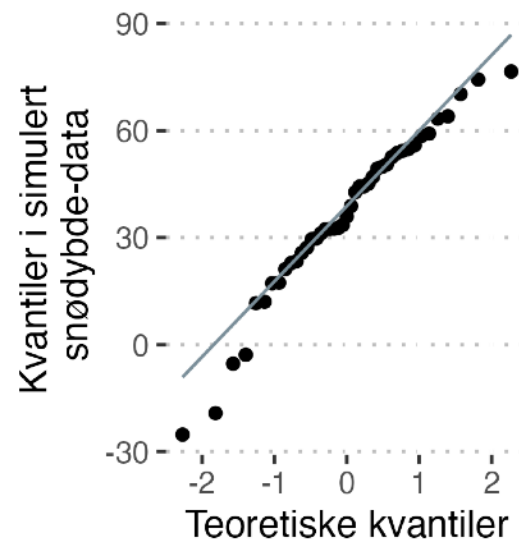
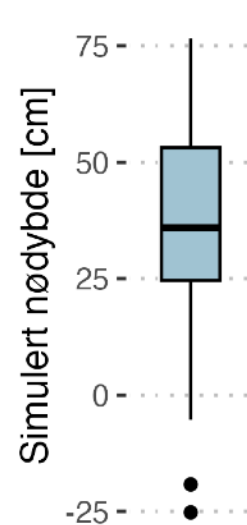
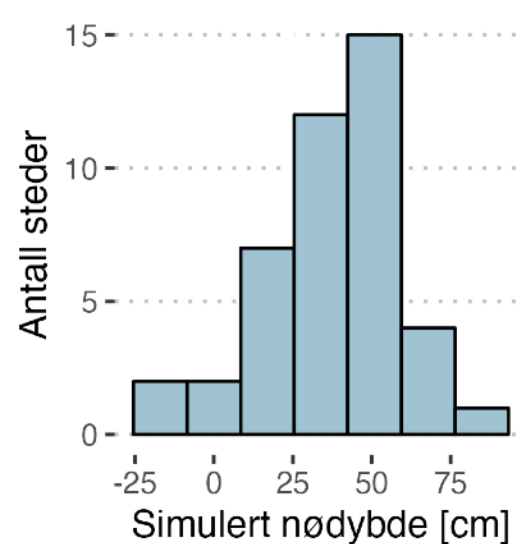
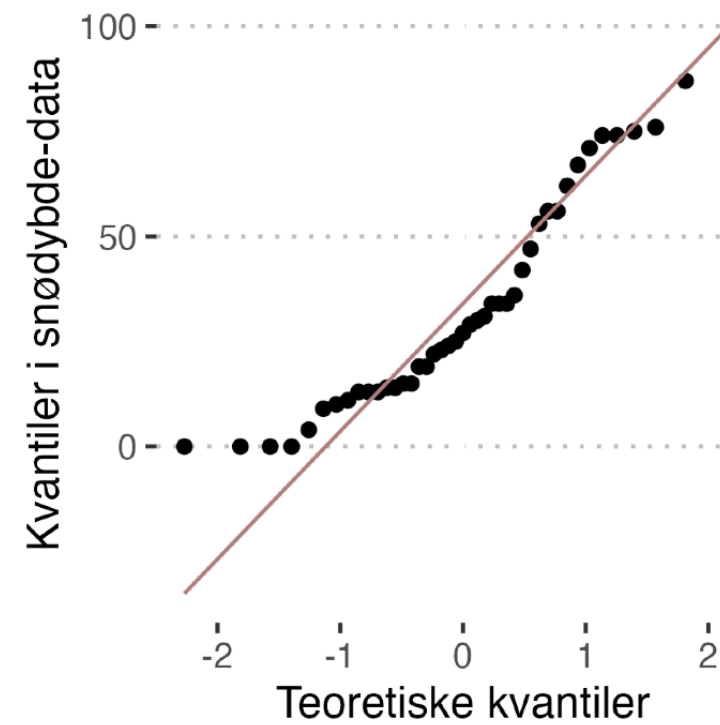
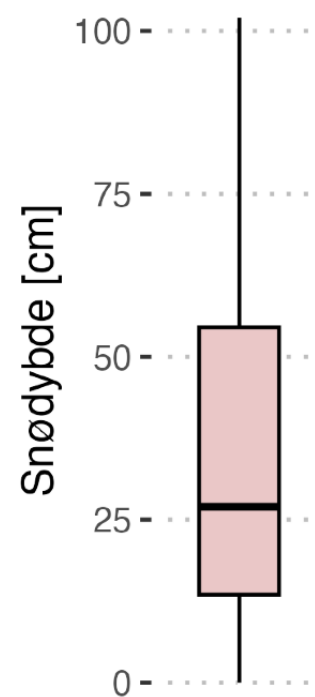
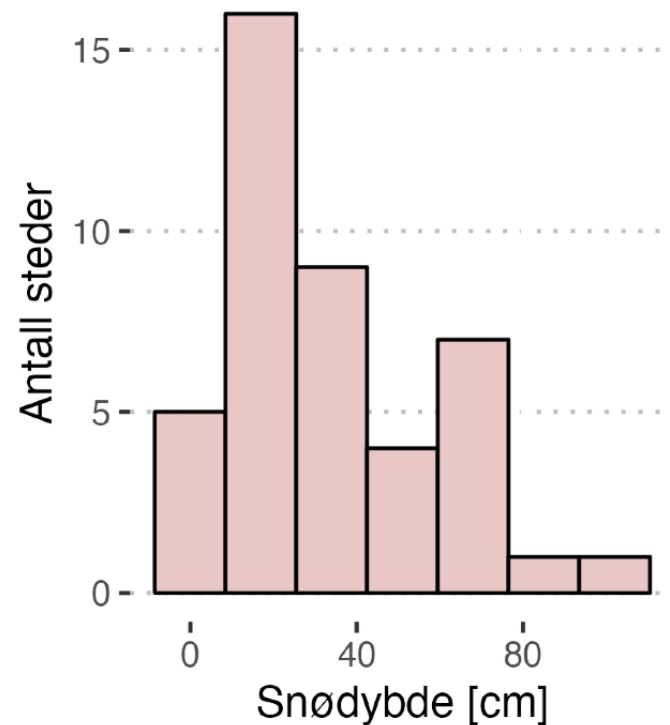
Spredningsplott

Snødybder i Trøndelag 22. februar 2024

Består av målinger av snødybde (i cm) fra 43 steder i Trøndelag

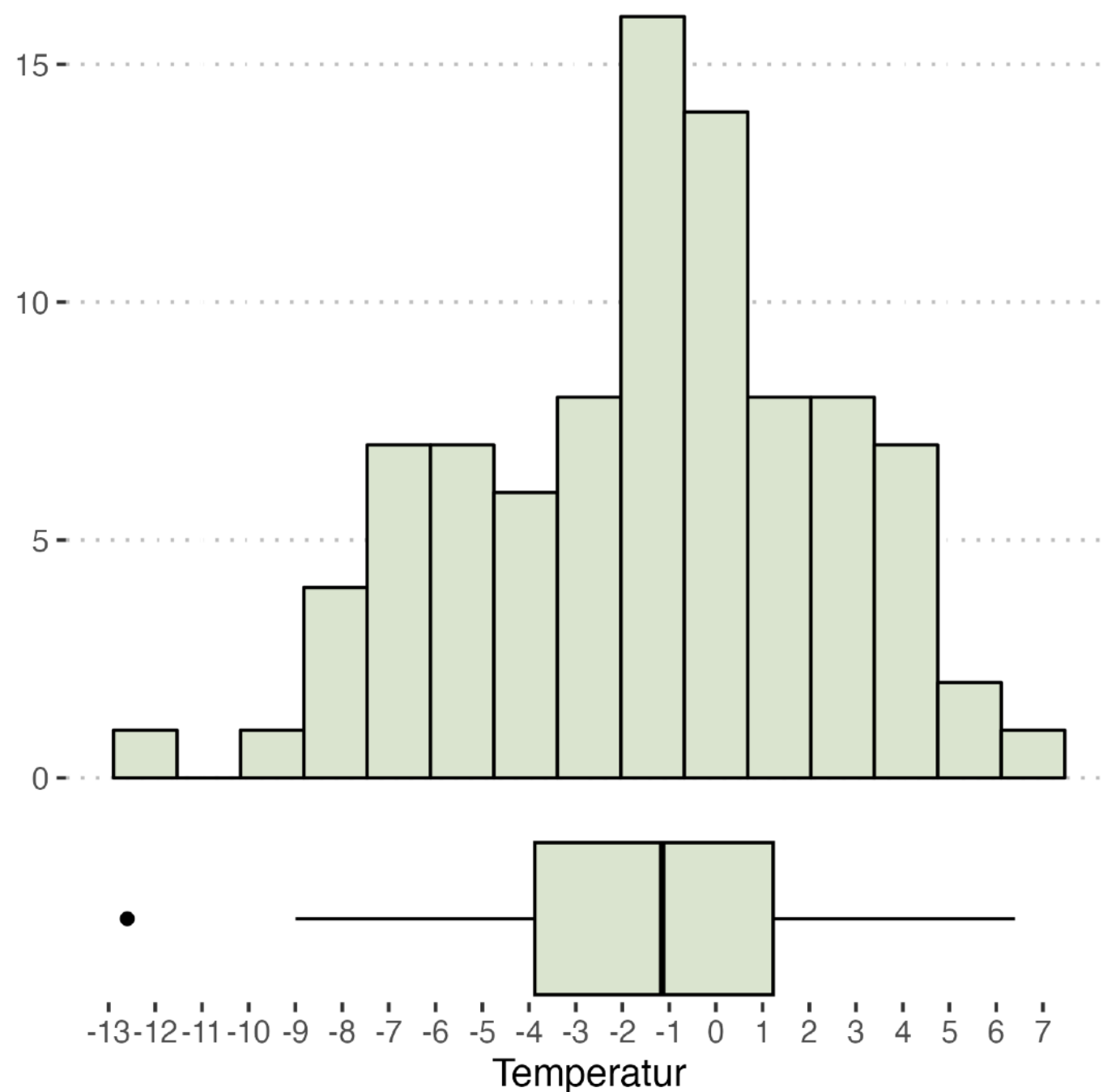


Hva er gjennomsnitt, median, standardavvik og kvartilbredde?



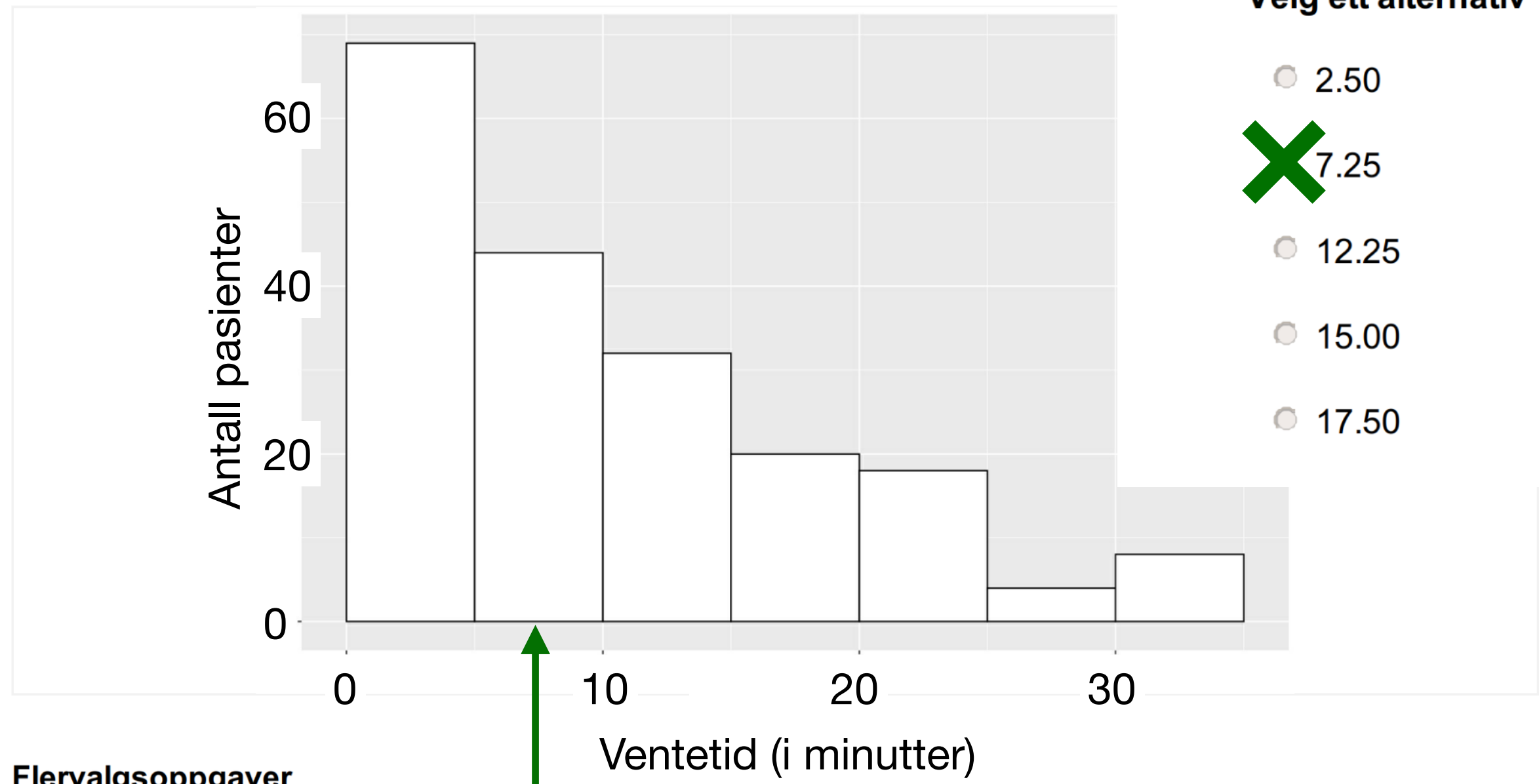
Temperatur i Trondheim

Består av målinger av temperatur (i grader Celsius) fra i januar, februar og mars 2023 i Trondheim



Hva er gjennomsnitt, median, standardavvik og kvartilbredde?

1(a)



Innledning: Histogrammet over viser frekvensfordelingen for ventetider for **200** pasienter på et tannlegekontor.

Oppgave: Hva er (omtrentlig) medianen av disse ventetidene?

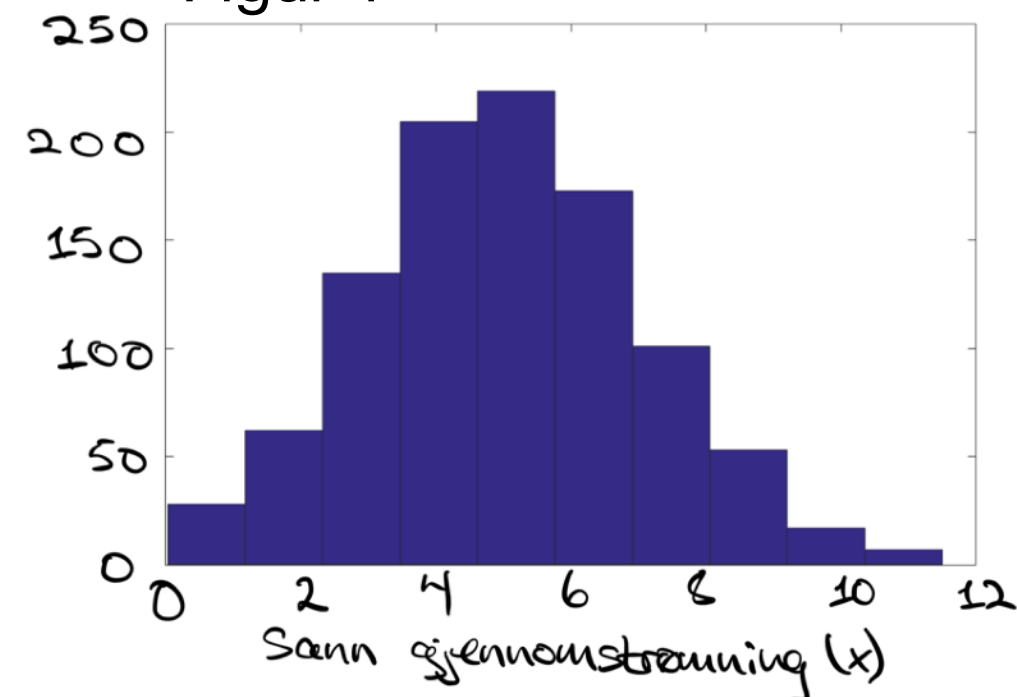
Oppgave 2 Den nye sensoren

Firmaet SkaffData prøver ut en ny sensor som skal gi billige data på gjennomstrømning av vann i rør. De prøver ut sensoren i en realistisk situasjon. I tillegg til målingene sensoren gir, $(y_1, y_2, \dots, y_n)$, måler de også tilhørende sann gjennomstrømning, $(x_1, x_2, \dots, x_n)$, for $n = 1000$ uavhengige tidsperioder. I figur 1 er histogrammet over sann gjennomstrømning, og i figur 2 er sann gjennomstrømning (x) plottet mot sensormålt gjennomstrømning (y).

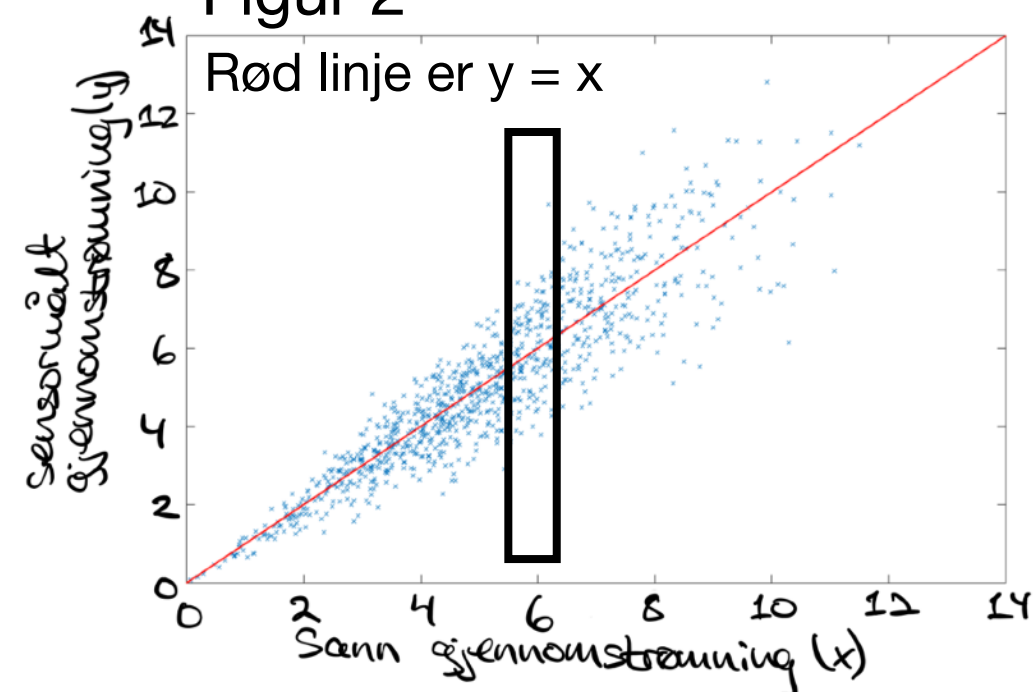
a) Basert på figur 1 og 2 svar på følgende spørsmål, og begrunn alle svarene kort:

- Hva er forventningsverdi og standardavvik til sann gjennomstrømning (X)? (Både forventningsverdi og standardavvik er heltall)
- Hva er forventningsverdi og standardavvik til sensormålt gjennomstrømning (Y) gitt at sann gjennomstrømning er $X = 6$. (Både forventningsverdi og standardavvik er igjen heltall)
- Er korrelasjonen (og kovariansen) mellom sann gjennomstrømning (X) og sensormålt gjennomstrømning (Y) positiv, negativ eller omtrent null?

Figur 1



Figur 2



Definisjon (Observator)

En observator er en observerbar funksjon av en eller flere stokastiske variabler som utgjør et utvalg

Definisjon (Estimator)

Anta at vi har uavhengige stokastiske variabler $X_1, X_2, \dots, X_n$, der $X_i \sim f_{X_i}(x_i; \theta)$ for $i = 1, 2, \dots, n$ og verdien til parameteren θ er ukjent. En estimator for θ er da en observator som benyttes til å anslå verdien til θ .

Eksempler på observatorer

- gjennomsnitt: $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
- median: $\tilde{X}$
- empirisk varians: $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$
- empirisk standardavvik: $S = \sqrt{S^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$

Oppgave 3 fra forrige onsdag

Frøya vil måle lengden (i cm) på en skjøteledning så nøyaktig som mulig. La μ betegne den ukjente lengden. Frøya gjør 5 uavhengige målinger med en meterstokk. Vi kan anta at hver måling X_i ($i = 1, \dots, 5$) er normalfordelt med forventningsverdi μ cm og standardavvik 1 cm. Senere får Frøya låne en lasermåler som hun bruker til å gjøre én måling. Vi kan anta at Y er normalfordelt med forventningsverdi μ cm og standardavvik 0.2 cm, og at Y er uavhengig av tidligere målinger.

Funksjoner av X_i og Y

For å anslå lengden μ vurderer Frøya fem ulike metoder:

$$\hat{\mu}_1 = Y, \hat{\mu}_2 = \frac{\sum_{i=1}^5 X_i}{5}, \hat{\mu}_3 = \frac{1}{2} \left(Y + \frac{\sum_{i=1}^5 X_i}{5} \right), \hat{\mu}_4 = \frac{Y + \sum_{i=1}^5 X_i}{6} \text{ og } \hat{\mu}_5 = 5Y - \sum_{i=1}^5 X_i$$

- a) Basert på forventningsverdi, hvilke(n) metode(r) er brukbar(e)?
- b) Basert på varians, hvilken metode er nyttigst?
- c) Hvilken metode er best?

“Metodene” er estimatorer.

$$\begin{aligned}\hat{\mu}_1 &= Y, \\ \hat{\mu}_2 &= \frac{\sum_{i=1}^5 X_i}{5}, \\ \hat{\mu}_3 &= \frac{1}{2} \left(Y + \frac{\sum_{i=1}^5 X_i}{5} \right), \\ \hat{\mu}_4 &= \frac{Y + \sum_{i=1}^5 X_i}{6} \text{ og} \\ \hat{\mu}_5 &= 5Y - \sum_{i=1}^5 X_i\end{aligned}$$

$$\begin{aligned}E(\hat{\mu}_1) &= \mu & \text{Var}(\hat{\mu}_1) &= 0.04 \\ E(\hat{\mu}_2) &= \mu & \text{Var}(\hat{\mu}_2) &= 0.2 \\ E(\hat{\mu}_3) &= \mu & \text{Var}(\hat{\mu}_3) &= 0.06 \\ E(\hat{\mu}_4) &= \mu & \text{Var}(\hat{\mu}_4) &= 0.14 \\ E(\hat{\mu}_5) &= 0 & \text{Var}(\hat{\mu}_5) &= 6\end{aligned}$$

Definisjon (Forventningsrett)

En estimator $\hat{\theta}$ sies å være en forventningsrett estimator for en parameter θ hvis

$$E[\hat{\theta}] = \theta.$$

Hvis dette ikke er oppfylt sies $\hat{\theta}$ å være en forventningsskjev estimator for θ .

Utelukker $\hat{\mu}_5$

Definisjon (Mest effisient estimator)

Av flere forventningsrette estimators for en parameter θ sier vi at den med minst varians er mest effisient.

Gir at $\hat{\mu}_1$ er best

Oppgave 3 Besøk på webside

For firmaet SeMeg er antall besøk på deres webside viktig. La X_i være antall besøk i løpet av t_i timer, og la $X_1, X_2, \dots, X_n$ være antall besøk i n ikke-overlappende tidsintervall. Vi antar at besøk på websiden er en poissonprosess med besøksintensitet λ . Dermed er $X_1, X_2, \dots, X_n$ uavhengige poissonfordelte stokastiske variabler med sannsynlighetsfordeling

$$f(x_i) = \frac{(\lambda t_i)^{x_i}}{x_i!} e^{-\lambda t_i} \quad \text{for } x_i = 0, 1, 2, \dots$$

a) Anta (bare i dette punktet) at $\lambda = 10$ og $t_1 = 1$. Finn sannsynlighetene

$$P(X_1 = 8) \quad , \quad P(X_1 \geq 8) \quad \text{og} \quad P(8 \leq X_1 \leq 12).$$

Vi antar nå at besøksintensiteten λ er ukjent. SeMeg ønsker å estimere intensiteten λ fra data på antall besøk fra n ikke-overlappende tidsintervall. Det er foreslått tre estimatorer,

$$\tilde{\lambda} = \frac{1}{n} \sum_{i=1}^n X_i \quad , \quad \hat{\lambda} = \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n t_i} \quad \text{og} \quad \hat{\hat{\lambda}} = \frac{1}{n} \sum_{i=1}^n \frac{X_i}{t_i},$$

og vi oppgir at $E[\hat{\lambda}] = \lambda$ og $\text{Var}[\hat{\lambda}] = \lambda / \sum_{i=1}^n t_i$.

b) Hvilken av de tre estimatorene vil du foretrekke når $n = 5$ og $t_1 = 1, t_2 = 2, t_3 = 5, t_4 = 1, t_5 = 5$? Begrunn svaret.

4(a) Binomisk fordelte variabler

Innledning: Anta at vi har to uavhengige stokastiske variabler X og Y , der $X \sim b(x; n, p)$ og $Y \sim b(y; n, 2p)$. X er altså antall suksesser i n uavhengige forsøk der hvert forsøk har sannsynlighet p for å gi suksess, mens Y er antall suksesser i n andre uavhengige forsøk der hvert forsøk gir suksess med sannsynlighet $2p$.

Oppgave:

- Når $n = 12$ og $p = 0.2$, finn følgende sannsynligheter,

$$P(X \leq 3), \quad P(Y \geq 4 | X \leq 3) \quad \text{og} \quad P(X + Y \leq 1).$$

A og B uavh.:

$$P(A \cap B) = P(A)P(B)$$

n	x	p											
		.05	.10	.15	.20	.25	.30	.40	.50	.60	.70	.80	.90
12	0	.540	.282	.142	.069	.032	.014	.002	.000	.000	.000	.000	.000
	1	.882	.659	.443	.275	.158	.085	.020	.003	.000	.000	.000	.000
	2	.980	.889	.736	.558	.391	.253	.083	.019	.003	.000	.000	.000
	3	.998	.974	.908	.795	.649	.493	.225	.073	.015	.002	.000	.000
	4	1.000	.996	.976	.927	.842	.724	.438	.194	.057	.009	.001	.000
	5	1.000	.999	.995	.981	.946	.882	.665	.387	.158	.039	.004	.000
	6	1.000	1.000	.999	.996	.986	.961	.842	.613	.335	.118	.019	.001
	7	1.000	1.000	1.000	.999	.997	.991	.943	.806	.562	.276	.073	.004
	8	1.000	1.000	1.000	1.000	1.000	.998	.985	.927	.775	.507	.205	.026
	9	1.000	1.000	1.000	1.000	1.000	1.000	.997	.981	.917	.747	.442	.111
	10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.997	.980	.915	.725	.341
	11	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.998	.986	.931	.718

4(a) Binomisk fordelte variabler

Innledning: Anta at vi har to uavhengige stokastiske variabler X og Y , der $X \sim b(x; n, p)$ og $Y \sim b(y; n, 2p)$. X er altså antall suksesser i n uavhengige forsøk der hvert forsøk har sannsynlighet p for å gi suksess, mens Y er antall suksesser i n andre uavhengige forsøk der hvert forsøk gir suksess med sannsynlighet $2p$.

Oppgave:

- Når $n = 12$ og $p = 0.2$, finn følgende sannsynligheter,

$$P(X \leq 3), \quad P(Y \geq 4 | X \leq 3) \quad \text{og} \quad P(X + Y \leq 1).$$

4(b) Innledning: I resten av oppgaven skal vi anta at verdien av parameteren $p \in [0, 0.5]$ er ukjent og skal estimeres basert på X og Y . Følgende tre estimatore er foreslått,

$$\hat{p} = \frac{X+Y}{2n}, \quad \tilde{p} = \frac{X+Y}{3n} \quad \text{og} \quad p^* = \frac{X}{2n} + \frac{Y}{4n} \quad (1)$$

Oppgave:

- Hvilken av de tre estimatorene vil du foretrekke? Begrunn svaret.

Binomisk fordeling

$$f(x) = b(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \dots, n.$$

$$E(X) = np, \quad \text{Var}(X) = np(1-p), \quad M_X(t) = (pe^t + 1 - p)^n.$$

Thomas har arvet en gullbarre etter besteforeldrene sine som han ønsker å selge. Gullbarren er av rent gull og veier μ gram. Før Thomas selger gullbarren ønsker han å sikre seg at han selger den for en korrekt pris og bestemmer seg derfor for å måle vekten til gullbaren på to måter.

Først bruker Thomas sin egen kjøkkenvekt til å veie gullbaren. Anta at vekten målt på kjøkkenvekten X er en normalfordelt stokastisk variabel med forventningsverdi μ og standardavvik 1 gram. Deretter går han til en forhandler for å måle vekten profesjonelt. La Y være vekten målt hos forhandleren, og anta at Y er normalfordelt med forventningsverdi μ og standardavvik 0.5 gram. Anta at $\text{Cov}(X, Y) = -0.2$.

For å estimere den sanne vekten μ gram til gullbarren vil Thomas sammenligne to ulike estimatorer

$$\hat{\mu} = Y \quad \text{og} \quad \tilde{\mu} = \frac{1}{2}(X + Y).$$

a) Beskriv kort hva som kjennetegner en god estimator.

Hvilken av de to estimatorene $\hat{\mu}$ og $\tilde{\mu}$ bør Thomas velge? Grunngi svaret ditt.

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y]$$