

Hjemmeeksamen i

MAT110 Statistikk 1

Eksamensdag	: Mandag 18. mai 2020
Tid	: 09:00 – 13:00 (4 timer) + 2 timer ekstra pga. det tekniske med Inspera
Hjelpemidler	: Alle trykte og skrevne hjelpemidler + kalkulator som kan inneholde data
Antall sider inkl. forsiden	: 25
Målform	: Norsk (bokmål)

- Det er ikke lov å få hjelp av andre.
- Dere skal jobbe alene.
- Faglærere kommer til å ta manuell plagiatkontroll.
- Reglene pålegger faglærerne å håndtere ulovlig samarbeid, hjelp eller kopiering som fusk.

Oppgave 1: (25 % - sannsynlighetsteori - koronatest)

I denne oppgaven skal vi se litt nærmere på personer som blir testet for koronavirus. Figur 1 viser de topp 5 landene over antall registrerte smittede per 8. april 2020.

Legg merke til følgende raritet:

[Tyskland](#), som har over 100 000 registrerte smittede personer, har *langt* lavere antall døde enn de andre landene på lista.

Region	↓ Registrert smittet	Per 100k innb.	Døde	Per 100k innb.
USA	399 929	119,7	12 925	3,9
Spania	146 690	313,7	14 555	31,1
Italia	135 586	224,3	17 129	28,3
Frankrike	110 070	161,6	10 343	15,2
Tyskland	107 663	128,5	2 016	2,4

Figur 1: Topp 5 landene i verden over antall registrerte smittede per. 8. april 2020. Kilde: vg.no/spesial/2020/corona/verden.

Årsaken for de lave dødelighetstallene i Tyskland kan være mange, men et av områdene hvor Tyskland skiller seg ut ift. de andre landene, er når det gjelder *testing* for koronavirus. Tyskland tester *svært* mange personer hver dag.

Foto: Heiko Junge



Foto: Elaine Thompson



Figur 2: Koronatesting.

Eksperiment:

Anta at vi velger en vilkårlig person i Tyskland og denne personen tar koronatesten. Vi registrerer deretter om testen gir positivt eller negativt svar. Til slutt finner vi ut om personen faktisk er syk eller ikke. ¹

Når en person testes for korona har man 4 mulige utfall: ²

Personen tester **positivt** for korona og er **smittet**. (korrekt test) (1)

Personen tester **positivt** for korona og er **ikke smittet**. (falsk positiv test) (2)

Personen tester **negativt** for korona og er **ikke smittet**. (korrekt test) (3)

Personen tester **negativt** for korona og er **smittet**. (falsk negativ test) (4)

Notasjon:

t = test er positiv (5)

$\bar{t}$ = test er negativ (6)

s = **smittet** person (7)

$\bar{s}$ = **ikke smittet** person (8)

Utfallsrom Ω :

$$\Omega = \left\{ (t, s) , (\bar{t}, s) , (t, \bar{s}) , (\bar{t}, \bar{s}) \right\} \quad (9)$$

a) **2.5%** Tolk sannsynlighetene $P(t, s)$, $P(\bar{t}, s)$, $P(t, \bar{s})$ og $P(\bar{t}, \bar{s})$ tilhørende utfallene til Ω , dvs. beskriv kort med ord hva sannsynlighetene betyr. ³

¹Vi kan finne ut om personen faktisk er smittet ved å ta enda en test som er enda sikrere, evt. vente lenge nok til at vi er sikre på at personen er syk eller ikke.

²En test er positiv dersom testen indikerer at man er smittet (men man må ikke være smittet selv om testen sier det).

³Tips: Bruk lign.(1)-(4).

Definerer begivenheter:

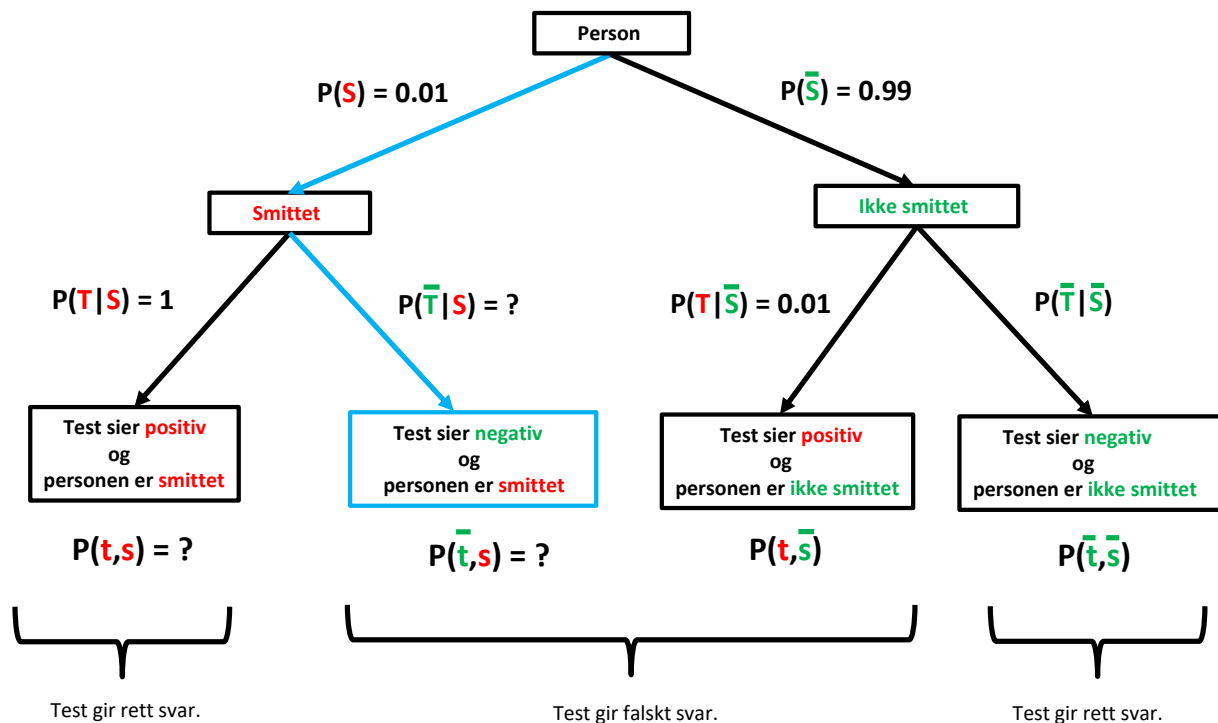
$$T = \{ (t, s), (t, \bar{s}) \} \quad (\text{testen er positiv}) \quad (10)$$

$$S = \{ (t, s), (\bar{t}, s) \} \quad (\text{personen er smittet}) \quad (11)$$

Anta at 1% av Tysklands befolkning er smittet, dvs.:

$$P(S) = 0.01 \quad (12)$$

Sannsynligheten oppgitt ovenfor samt andre relevante sannsynligheter tilhørende de 4 mulige utfallene i utfallsrommet Ω i lign.(9) kan visualiseres i et sannsynlighetstre:



Figur 3: Sannsynlighetstre.

Egenskaper til test:

Det finnes flere forskjellige typer koronatester.

En test designes alltid slik at sannsynligheten for å få en falsk negativ test er svært lav, dvs. at testen gir negativt resultat selv om personen er smittet av korona (utfall i lign.(4)).

Dette går som regel på bekostning av den andre måten en test kan gå galt, nemlig falsk positiv, dvs. at testen gir positivt resultat selv om personen *ikke* er smittet av korona (utfall i lign.(2)).

Vi skal nå regne på hvordan selv en relativt lav sannsynlighet for falsk positiv test kan forklare noe av hvorfor Tyskland har så lave dødelighetstall.

La oss i denne oppgaven se på en test som har følgende egenskaper:

- 1) Anta at dersom en person som er **smittet** tar testen så blir testresultatet **positivt** med 100 % sannsynlighet. Matematisk kan dette formuleres slik:

$$P(T | S) = 1 \quad (13)$$

Altså koronatesten gir alltid rett resultat dersom testen utføres på personer som er smittet.

- 2) Anta at dersom en person som **ikke** er **smittet** tar testen så blir testresultatet også **positivt** med 1 % sannsynlighet. Matematisk kan dette formuleres slik:

$$P(T | \overline{S}) = 0.01 \quad (14)$$

Altså koronatesten gir feil resultat i 1 % av tilfellene dersom testen utføres på en person som er frisk, dvs. ikke er smittet.

De betingede sannsynlighetene i lign.(13) og (14) er vist i sannsynlighetstreet i figur 3.

b) 5% Regn ut sannsynlighetene $P(\overline{T} | S)$ og $P(\overline{t}, s)$.⁴

c) 2.5% Hva sier svaret til $P(\overline{t}, s)$ fra oppgave **1b)** om falske negative tester?⁵

d) 5% Bruk setningen om oppsplitting av utfallsrom og vis ved regning at sannsynlighet for at testen gir positivt resultat $P(T)$ er

$$P(T) = 0.0199 \quad (15)$$

e) 5% Bruk svaret fra oppgave **1d)**, dvs. lign.(15), og vis ved regning at

$$P(\overline{S} | T) = 0.4975 \quad (16)$$

f) 5% I lys av svaret fra oppgave **1e)**, dvs. lign.(16), er det lurt å teste friske personer? Begrunn svaret.⁶



⁴Ta gjerne en titt på sannsynlighetstreet. Bruk læresetninger som oppgitt i kompendiet når dere regner ut svarene, f.eks. komplementsetningen og multiplikasjonssetningen.

⁵Kun et kort svar trengs her.

⁶Kun et kort svar trengs her.

Oppgave 2: (25 % - stokastiske variabler og normalfordeling - aldersfordeling)

Per 16. april 2020 var det registrert totalt $N = 6791$ personer som testet positivt på korona i Norge. Disse personene ble delt inn i 10 grupper etter alder, 0-9 år, 10-19 år osv. og opp til 90-99 år.

Aldersfordelingen til disse aldersgruppene er vist i figur 4:

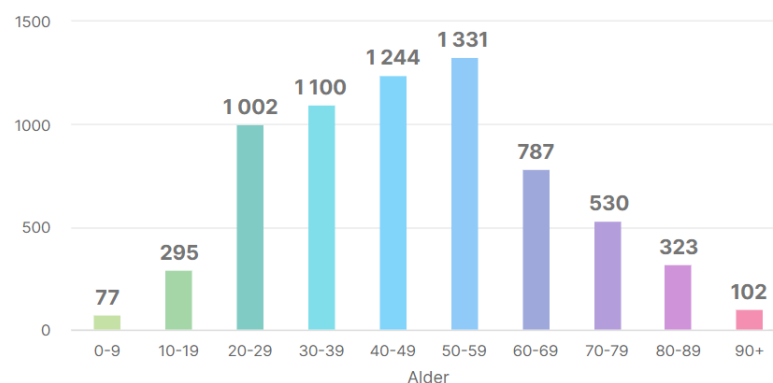
Relativfrekvensene tilhørende disse aldersgruppene kaller vi $p_1, p_2, \dots, p_{10}$, se tabell 1.

	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}
Relativfrekvens p_i	$\frac{77}{6791}$	$\frac{295}{6791}$	$\frac{1002}{6791}$	$\frac{1100}{6791}$	$\frac{1244}{6791}$	$\frac{1331}{6791}$	$\frac{787}{6791}$	$\frac{530}{6791}$	$\frac{323}{6791}$	$\frac{102}{6791}$
Aldersgruppe	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90+

Tabell 1: Relativfrekvenser p_i , hvor $i = 1..10$, fra figur 4.

Aldersfordeling

Tallene viser aldersfordelingen blant de som har testet positivt. Tallene oppdateres daglig fra FHIs dagsrapport og kommer kun på landsnivå.



Sist oppdatert: 16. april (i dag).

Figur 4: Aldersfordeling blant de $N = 6791$ personene som testet positivt per. 16. april 2020.
Kilde: vg.no/spesial/2020/corona.

Eksperiment E :

1. Vi velger en vilkårlig person av alle $N = 6791$ personer som har testet positivt på korona.
2. Vi registrerer hvilken aldersgruppe den valgte personen tilhører.
3. Utfallet til eksperimentet er gjennomsnittsalderen for den aktuelle aldersgruppen. ⁷

Diskret stokastisk variabel X : (basert på eksperimentet E)

$$X = \text{gjennomsnittsalderen av aldersgruppen som en tilfeldig valgt person tilhører} \quad (17)$$

som har testet positivt for korona

Utfallsrommet til X :

$$\Omega = \left\{ 4.5, 14.5, 24.5, 34.5, 44.5, 54.5, 64.5, 74.5, 84.5, 94.5 \right\} \quad (18)$$

Anta:

$$P(X = x_i) = p_i \quad (19)$$

hvor x_i er gjennomsnittsalderen for aldersgruppen i , med $i = 1, \dots, 10$.

a) 7.5 % Vis, ved regning, at forventningen og variansen til X er: ⁸ ⁹

$$E[X] = 47.45 \quad (\text{år}) \quad (21)$$

$$Var[X] = 373.16 \quad (\text{år}^2) \quad (22)$$

Regn også ut tilhørende standardavvik $\sigma[X]$.

⁷F.eks. dersom personen tilhører aldersgruppe 0-9 år så blir utfallet $\frac{0+9}{2} = 4.5$ år.

⁸Bruk definisjonen av forventning til å regne ut $E[X]$.

⁹Anta

$$E[X^2] = 2624.66 \quad (\text{år}^2) \quad (20)$$

og bruk “varianssetningen” til å regne ut $Var[X]$ i lign.(22).

Eksperiment E' :

Anta at vi gjennomfører et nytt eksperiment E' hvor vi igjen velger en tilfeldig person som har testet positivt for korona, men at utfallet nå er den faktiske alderen til den valgte personen.

Kontinuerlig stokastisk variabel Y : (basert på eksperimentet E')

$$Y = \text{alder til en tilfeldig utvalgt person som har testet positivt for korona} \quad (23)$$

i eksperiment E'

Anta:

$$Y \sim N[E[X], \sigma[X]] \quad (24)$$

dvs. Y er normalfordelt med $E[Y] = E[X]$ og $\sigma[Y] = \sigma[X]$ gitt ved verdiene fra oppgave 2a. Figur 5 viser vi hvordan dataene tilpasses en normalfordeling.

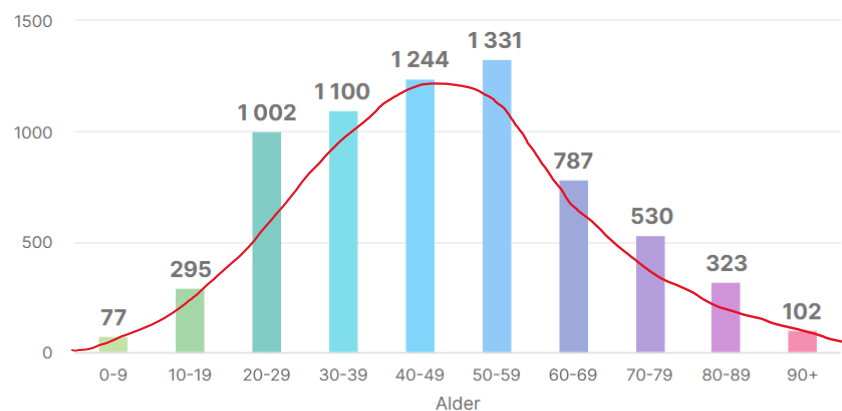
b) 2.5% Vis ved regning at

$$P(Y < 20) = 0.0778 \quad (25)$$

Forklar kort med ord hva denne sannsynligheten sier.

Aldersfordeling

Tallene viser aldersfordelingen blant de som har testet positivt. Tallene oppdateres daglig fra FHIs dagsrapport og kommer kun på landsnivå.



Sist oppdatert: 16. april (i dag).

Figur 5: Aldersfordelingen Y antas å være normalfordelt.

Legg merke til at det er et uvanlig stort *gap* mellom aldersgruppene 10-19 og 20-29, se figur 6.

En av grunnene til det dette høye gapet kan være et uvanlig stort *mørketall* m for aldersgruppen 10-19.¹⁰

Her er noen mulige grunner for hvorfor akkurat aldersgruppen 10-19 har høyere mørketall enn de andre aldersgruppene:

- Barn og ungdom som er smittet av korona har ofte milde symptomer, se figur 7.

Fra forskning.no:

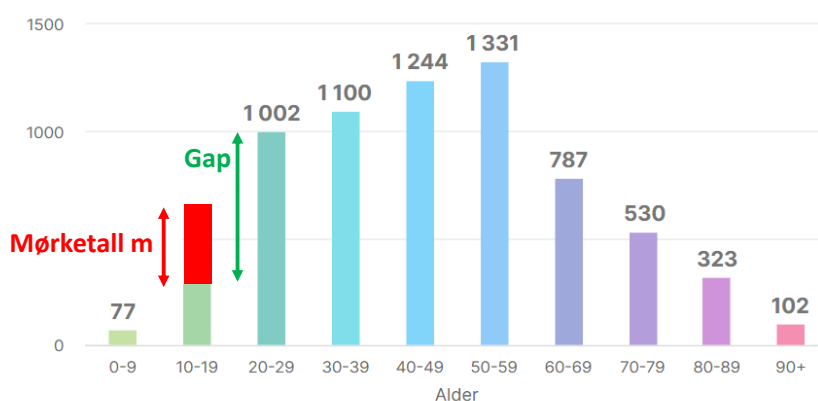
“Barn rammes trolig sjeldnere og de får klart mildere symptomer av viruset.”

- Barn og ungdom har ofte betydelig større kontakthypighet med andre mennesker. Av disse kan man si at aldersgruppen 10-19 har ennå større kontakthypighet enn 0-9 år og har dermed størst potensial for mørketall.

Vi skal videre i denne oppgaven forsøke å finne mørketallet m og bruke dette til å korrigere normalfordelingen.

Aldersfordeling

Tallene viser aldersfordelingen blant de som har testet positivt. Tallene oppdateres daglig fra FHIs dagsrapport og kommer kun på landsnivå.



Sist oppdatert: 16. april (i dag).

Figur 6: Mørketallet m og gapet mellom aldersgruppene 10-19 og 20-29.

¹⁰Med *mørketall* mener vi her personer som har blitt smittet av korona, men som ikke har blitt registrert.



Ikke ett eneste barn under ti år har til nå dødd av koronaviruset. Barn rammes trolig sjeldnere og de får klart mildere symptomer av viruset. Dette er så langt et mysterium. Finner forskere svaret, kan det hjelpe oss i kampen mot viruset. (Montasje: Shutterstock/NTB scanpix/forskning.no)

Hvorfor rammes ikke barn like hardt av koronaviruset?

Figur 7: Artikkel i *forskning.no* fra 20. mars 2020.

La $E'[X]$ være den *korrigerte* forventningsverdien for Y dersom vi hadde inkludert mørketallet m for aldersgruppen 10-19.

- c) 5% Dersom den “korrigerte” forventningen $E'[X]$ forklares av mørketallet m for aldersgruppen 10-19 år, vis ved regning at:

$$E'[X] = \frac{E[X] N + 14.5m}{N + m} \quad (26)$$

Hvordan en sykdom sprer seg ift. alder er ofte avhengig av visse karakteristikk ved sykdommen. Sammenlignet med en annen sykdom med tilsvarende karakteristikk som korona, var gjennomsnittsalderen for en smittet person for denne sykdommen 46 år.

La oss derfor anta at den korrigerte forventningsverdien $E'[X]$ har samme tall som den lignende sykdommen, dvs. anta at:

$$E'[X] = 46 \quad (\text{år}) \quad (27)$$

- d) 5% Regn ut tallverdien til mørketallet m under antagelsen i lign.(27).
ved å bruke lign.(26).¹¹

Når vi nå har regnet ut mørketallet m , kan vi enkelt regne ut korrigert standardavvik:¹²

$$\sigma'[X] = 20.07 \quad (28)$$

- e) 5% Med denne nye modellen, regn ut nå $P(Y < 20)$.

Forventer du nå større eller mindre sannsynlighet for $P(Y < 20)$ sammenlignet med oppgave 2b? Gi en kort begrunnelse.¹³

■

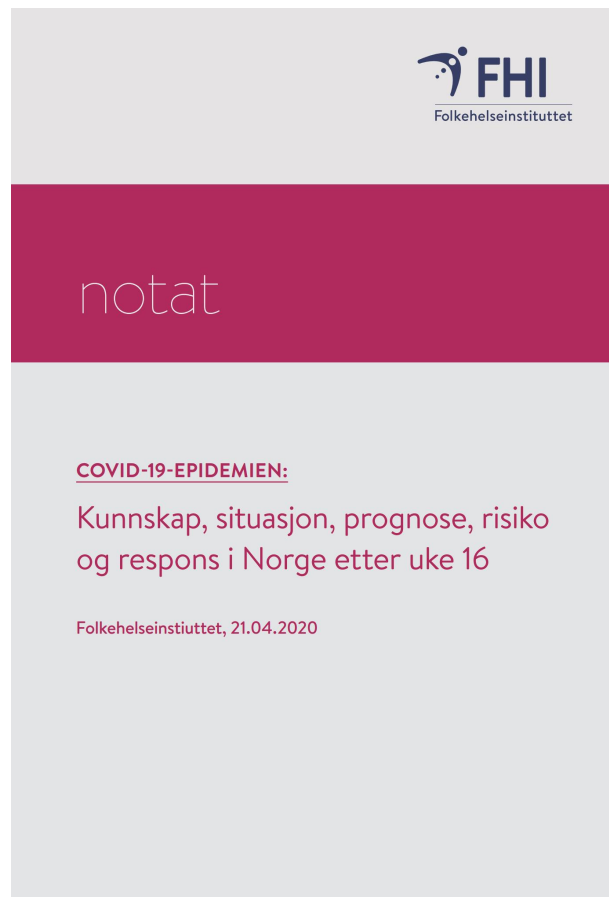
¹¹Rund av til nærmeste heltall.

¹²Du skal ikke regne ut lign.(28). Bare ta den for gitt.

¹³Du kan enten gi din begrunnelse via en kort regning, eller du kan argumentere. Begrunnelse via regning er muligens enklest og mest presist.

Oppgave 3: (25% - statistisk modellering - bremse koronaspredning)

Denne oppgaven er basert på et notat, publisert 21. april 2020, skrevet av Folkehelseinstituttet (FHI) og ble publisert 21. april 2020, se figur 8:



Figur 8: Notat utgitt av FHI.

Link til dokumentet finnes her.

På side 40 i notatet analyseres hvordan responsen av diverse tiltak har på det såkalte *reproduksjonstallet*.

I figur 9 har vi oppgitt et utdrag av side 40 i notatet.

10. Respons

Epidemien herjer videre i verden og vil gjøre det i flere år. Den norske befolkningen er ikke immun. Vi må finne en farbar vei videre for befolkningen, helsetjenesten og samfunnet. Bare utbredt immunitet i befolkningen gir en varig løsning på epidemien.

10.1 Grunnlaget for smittevernstrategien

En smittevernstrategi består av et mål og et sett av tiltak som skal bringes oss til det målet. Strategien bør bygge på overvåking av epidemien, oppdatert kunnskap og nye analyser og vurderinger.

Tiltakene mot epidemiens spredning tar sikte på å redusere reproduksjonstallet (antallet nye tilfeller som en gjennomsnittspasient gir opphav til) fra det antatte utgangspunkt (R_0) på rundt 2,5 til et betydelig lavere effektivt tall (R_E).

Reproduksjonstallet bestemmes (noe forenklet) av fire faktorer:

$$R_E = \beta_e c_e D_e x$$

Smittsomheten β kan påvirkes ved hygienetiltak (tiltak 1) slik at den effektive (faktiske) smittsomheten blir lavere: $\beta_e < \beta$.

Kontakthypigheten c kan påvirkes med tiltak (tiltak 2-5) som skiller smittsomme personer fra mottakelige personer slik at den effektive (faktiske) kontakthypigheten blir lavere: $c_e < c$. Siden de smittsomme personene kan mangle symptomer, ha få symptomer eller ha symptomer som lett forveksles med andre sykdommer, er det ikke mulig å stoppe epidemien med bare tidlig oppdaging og isolering av de smitteførende. Det trengs karantene for nærkontakter og antakelig også tiltak som reduserer hypigheten av nærkontakt mellom alle mennesker siden det blant dem kan være ikke-erkjente smittsomme personer.

Varigheten D av smittsom periode kan påvirkes dersom det kommer et legemiddel (tiltak 6) som direkte angriper viruset og forkorter den tiden pasienten er smitteførende, $D_e < D$. (Noen vil velge å se på isolering av smitteførende som en forkorting av den effektive smitteførende perioden i stedet for redusert kontakthypighet.)

Andelen mottakelige x synker over tid som følge av at mange gjennomgår infeksjonen. Etter hvert blir x så stor at R_E faller under 1 og epidemien dør ut. Høy befolkningsimmunitet (lav x) kan gi en varig beskyttelse mot nye, store utbrudd. Dersom vaksinasjon (tiltak 7) blir tilgjengelig, kan x reduseres på kunstig vis.

Figur 9: Utdrag av side 40 i notatet fra FHI.

Det basale reproduksjonstallet $\mathbb{R}_0$:

Det basale reproduksjonstallet:

$$\mathbb{R}_0 = \frac{\text{antall direkte smittoverføringer en smittet person vil gi i en populasjon når andelen mottagelige personer er 1}}{\text{andelen mottagelige personer er 1}} \quad (29)$$

At andelen mottagelige personer er 1, betyr i prinsippet at ingen personer i populasjonen vi ser på ennå har blitt smittet. ¹⁴

Matematisk modell:

$$\mathbb{R}_0 = \beta c D \quad (30)$$

hvor ¹⁵

$$\beta = \text{sannsynligheten for smitteoverføring} \quad (31)$$

$$c = \text{kontakthypigheten} \quad (32)$$

$$D = \text{varigheten av smittsom periode} \quad (33)$$

Sannsynligheten β for smitteoverføring er sannsynligheten for at man faktisk *blir* smittet dersom man er i kontakt med en smittet person.

Kontakthypigheten c er forventet antall ganger en smittet person er i kontakt med person mottagelig for smitten *per tidsenhet*. ¹⁶

Varigheten D av en smittsom periode sier hvor lang tid en smittet person faktisk kan smitte andre personer. ¹⁷

Legg merke til at $\mathbb{R}_0$ er en *dimensjonsløs* (ingen enhet, bare et tall) og sammensatt størrelse som ikke bare gjenspeiler egenskaper ved *sykdommen* (β og D), men også egenskaper ved *populasjonen* (c). Derfor opererer man med forskjellig $\mathbb{R}_0$ for forskjellige populasjoner, f.eks. Oslo, som viser høy smittespredning har derfor høy $\mathbb{R}_0$, mens Møre og Romsdal som har lav smittespredning har lav $\mathbb{R}_0$. ¹⁸

¹⁴Den første personen som er smittet kalles ofte “pasient zero” eller “index zero”, og $\mathbb{R}_0$ sier ganske enkelt hvor mange direkte nye smittetilfeller “pasient zero” forårsaker.

¹⁵konstantene β , c og D er konstanter

¹⁶For eksempel, hvis $c = 2$ per dag, betyr dette at en smittet person forventes å være i kontakt med to mottagelige personer per dag.

¹⁷For eksempel, hvis $D = 10$ dager, betyr dette at en smittet person kan smitte andre personer i totalt 10 dager.

¹⁸Hvert land har f.eks. sin egen $\mathbb{R}_0$. FHI opplyser i notatet at *Norge* (som populasjon) har $\mathbb{R}_0 = 2.5$.

1. Anta at N personer er samlet på et avgrenset sted hvor ingen personer kommer til eller fra.
2. Deler de N personene inn i tre kategorier: ²⁰

$$\mathbf{S} = \text{Susceptibles} - \text{mottagelige} \quad (\text{personer som kan bli smittet}) \quad (34)$$

$$\mathbf{I} = \text{Infectious} - \text{smittett} \quad (\text{personer som kan smitte andre}) \quad (35)$$

$$\mathbf{R} = \text{Recovered} - \text{immune} \quad (\text{personer som er immune eller døde}) \quad (36)$$

3. I uke $t = 0$ har kun én av personene korona, nemlig pasient zero.
Anta at pasient zero er uvitende om at han/hun er smittet med korona.
Ingen av de $N - 1$ andre personene er immune mot korona.
4. En person som blir smittet i en gitt uke kan kun overføre smitten **den påfølgende uken**.
Da smitter personen hver smittemottagelig person med sannsynlighet p .
I løpet av denne uken blir personen immunt eller dør. ²¹
5. Før starten av hver uke registreres antall personer i hver kategori. ²²
6. Vi gjentar eksperimentet uke for uke inntil epidemien stopper opp, dvs. ingen nye smittetilfeller inntreffer.



Figur 10: Kupé-modell.

¹⁹Vi påpeker at dette er en sterkt forenklet modell for koronaviruset. Hensikten er å få frem prinsippene bak hvordan man modellerer smittespredning.

²⁰Dette kalles en SIR-modell, og er den enkleste såkalte “kupé-modellen” innen epidemiologi.

²¹Altså personen er frisk og smitter ingen andre mer.

²²Vi tester mao. alle personene som er smittemottagelige og oppdaterer tallene for kategoriene.

Eksempel $N = 3$:

Anta at vi har en populasjon på $N = 3$ personer:

- ei dame
- en mann
- pasient zero (første smittede person)

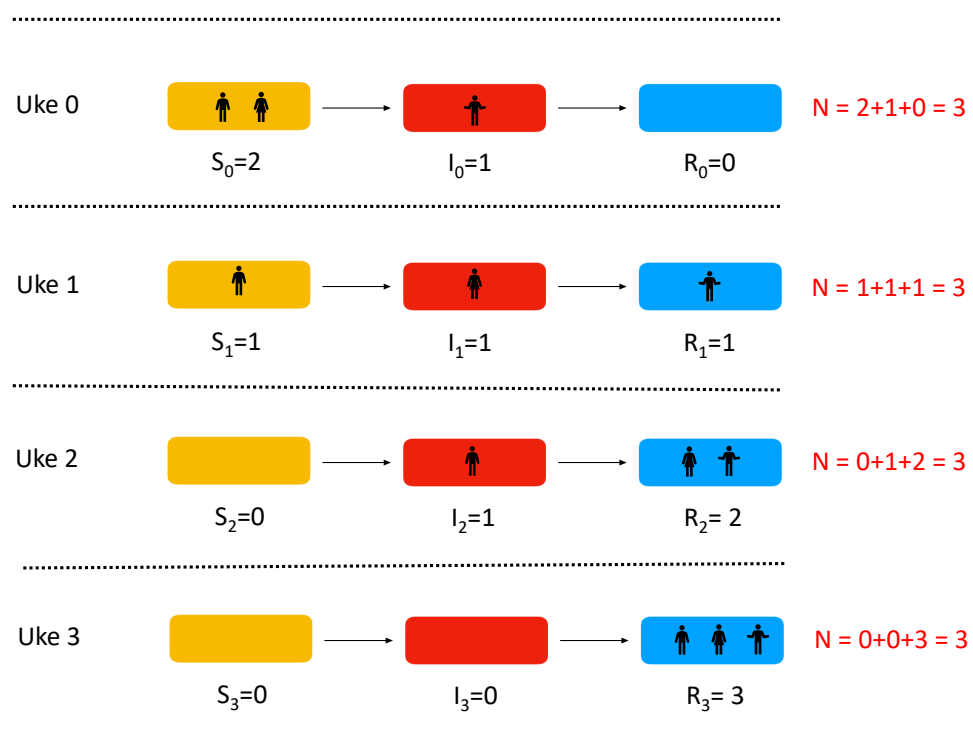
Figur 11 viser en mulig utvikling i tid for $N = 3$.

Uke 0 Pasient zero er smittet og de 2 andre er mottagelige for smitte.

Uke 1 Damen er smittet (av pasient zero), men ikke mannen.
Mannen er fremdeles mottagelig for smitte.
Pasient zero har blitt immun.

Uke 2 Mannen er smittet (av damen). Damen er immun.

Uke 3 Mannen er immun. Alle er immune. Ingen nye kan bli smittet.
Epidemien er slutt.



Figur 11: Et mulig forløp med $N = 3$.

Sannsynlighetsmodell:

Stokastiske variabler:

$$S_n = \text{antall personer som er mottagelige for smitte ved slutten av uke } n, \quad n = 0, 1, 2, 3 \dots \quad (37)$$

$$I_n = \text{antall smittede personer ved slutten av uke } n \quad n = 0, 1, 2, 3 \dots \quad (38)$$

$$R_n = \text{antall immune eller døde personer ved slutten av uke } n \quad n = 0, 1, 2, 3 \dots \quad (39)$$

hvor tilstanden for en uke n er beskrevet ved begivenheten B_n :

$$B_n = B_n(s_n, i_n, r_n) \quad (40)$$

$$= \{ S_n = s_n, I_n = i_n, R_n = r_n \} \quad n = 1, 2, 3 \dots \quad (41)$$

med startverdiene

$$S_0 = N - 1 \quad (42)$$

$$I_0 = 1 \quad (43)$$

$$R_0 = 0 \quad (44)$$

dvs.

$$B_0 = B_0(N - 1, 1, 0) \quad (45)$$

$$= \{ S_0 = 2, I_0 = 1, R_0 = 0 \} \quad (46)$$

Dersom vi *betinges* på SIR-verdiene ²³ ved slutten av denne uken, uke n , vil fordelingen for antall smittede ved slutten av neste uke I_{n+1} være **binomialfordelt** med sannsynlighet p^{i_n} , hvor $I_n = i_n > 0$:

$$I_{n+1} \mid B_n(s_n, i_n, r_n) \sim \text{Bin}[s_n, p^{i_n}] \quad n = 0, 1, 2, 3 \dots \quad (47)$$

For S_{n+1} og R_{n+1} har vi følgende sammenhenger:

$$S_{n+1} = S_n - I_{n+1} \quad n = 0, 1, 2, 3 \dots \quad (48)$$

$$R_{n+1} = R_n + I_n \quad n = 0, 1, 2, 3 \dots \quad (49)$$

²³Verdiene for S_n , I_n og R_n for en gitt uke kalles ofte "SIR-verdiene".

I tillegg til antagelsen i lign.(47), antar vi at prosessen er kun avhengig av tilstanden fra *forrige uke* og ikke ukene før der, dvs. “korttidshuholdelse”-antagelse: ²⁴

$$P(I_{n+1} = i_{n+1} \mid B_n, B_{n-1}, \dots, B_1, B_0) = P(I_{n+1} = i_{n+1} \mid B_n), \quad n = 0, 1, 2, \dots \quad (50)$$

a) **5 %** Med $N = 3$ er $B_0 = B_0(2, 1, 0)$. Vis da ved regning at: ²⁵

$$P(I_1 = 1 \mid B_0) = 2p - 2p^2 \quad (51)$$

b) **5 %** Med $N = 3$ er $B_0 = B_0(2, 1, 0)$. Vis da ved regning at: ²⁶

$$P(I_1 = 2 \mid B_0) = p^2 \quad (52)$$

c) **2.5 %** La nå $B_0 = B_0(N - 1, 1, 0)$.
Fra innledningen vet vi at forventet antall smittede etter dag 1 er lik $\mathbb{R}_0$,
se lign.(29) og (30), dvs. $E[I_1] = \mathbb{R}_0$. Bruk lign.(47) og vis at ²⁷

$$E[I_1 \mid B_0] = (N - 1)p \quad (53)$$

d) **2.5 %** I vår modell har vi satt smittsom periode til $D = 1$ uke, dvs. én periode.
Anta nå at:

$$\beta = 0.5 \quad (\text{sannsynlighet for å overføre sykdommen ved kontakt}) \quad (54)$$

$$c = 3 \quad (\text{forventet antall kontakter ila. en uke}) \quad (55)$$

Anta $N = 3$, bruk lign.(53) og regn ut verdien til p .

²⁴Denne sammenhengen ser mer skremmende ut enn den er, den sier bare at prosessen husker kun hva som skjedde én uke siden. Dette er en rimelig antagelse i vårt tilfelle, siden spredningen starter “på nytt” etter hver uke. Dette kalles en Markov tilnærming, “Markov approximation”.

²⁵Bruk lign.(47).

²⁶Bruk lign.(47).

²⁷Argumenter gjerne ved å bruke lign.(47). Husk at dersom $X \sim \text{Bin}[m, p]$ så er $E[X] = mp$.

I løpet av en epidemi er vi alltid interessert *to* størrelser:

- Det *totale* antallet personer som ble smittet.
- Det *maksimale* antallet personer som var smittet samtidig.

Stokastiske variabler:

$$Z = \text{totalt antalle personer som ble smittet ila. hele epidemien} \quad (56)$$

$$M = \text{maskimalt antall personer var smittet samtidig ila. hele epidemien} \quad (57)$$

Anta at vi ønsker å bruke *isolasjon* for å begrense spredningen av korona, dvs. begrense Z .

Å bruke isolasjon, betyr at vi *reduserer* kontakthypigheten c til en *lavere* verdi c' .

e) 10% Anta at vi har gitt tallene fra oppgave 3d og med startverdiene: ²⁸

$$B_0 = B_0(2, 1, 0) \quad (58)$$

$$= \{ S_0 = 2, I_0 = 1, R_0 = 0 \} \quad (59)$$

Finn ved regning hva den reduserte kontakthyppheten c' må være for at sannsynligheten for at hele populasjonen $N = 3$ blir smittet, dvs. $Z = 3$, er 0.01. Altså, finn c' slik at: ²⁹ ³⁰ ³¹

$$P(Z = 3 \mid B_0) = 0.01 \quad (63)$$

²⁸Dvs. anta at $N = 3$, $D = 1$, $c = 3$ og $\beta = \frac{1}{2}$.

²⁹Dette er en A-oppgave. Hopp gjerne over denne oppgaven og prøv mot slutten dersom tiden strekker til.

³⁰Bruk at

$$Z = I_1 + I_2 + I_3 + \dots \quad (60)$$

Bruk så gjerne figur 11 for å finne ut når epidemien dør ut, dvs i hvilken uke k vi har første gang at $I_k = 0$.

³¹Dersom du gjør oppgaven korrekt får du en 3. gradsligning på formen:

$$3p^2 - 2p^3 = 0.01 \quad (61)$$

Denne ligningen har *tre* mulige løsninger:

$$p = 1.4977, \quad p = -0.0567, \quad p = 0.0589 \quad (62)$$

Oppgave 4: (25% - regresjonsanalyse - antall smittede)

Figur 12 viser antall smittede av koronaviruset i Norge per 27. april 2020.

Dersom vi nå ser på perioden fra og med 13. mars (stiplet vertikal linje i figur 12) til og med 27. april, dvs. totalt 47 dager, så ser vi at grafen viser en tilnærmet lineær vekst.³²

Notasjon:

$$x_i = \text{dag nr. } i \text{ regnet fra 13. mars 2020, hvor } i = 0, \dots, 46 \quad (64)$$

$$y_i = \text{antall registrert smittede personer av koronaviruset i Norge dag nr. } i \quad (65)$$

dvs. dag nr. $i = 0$ er 13. mars og dag $i = 46$ er 27. april 2020.³³

Man kan regne ut gjennomsnittene $\bar{x}$ og $\bar{y}$:³⁴

$$\bar{x} = 23 \quad , \quad \bar{y} = 4959.34 \quad (66)$$



Figur 12: Antall registrerte smittede i Norge per 27. april 2020. Kilde: vg.no/spesial/2020/corona.

³²På datoen 13. mars 2020 ble det “*endring av testkriterier*”, dvs. kriteriene for hvem som skulle bli testet, ble endret.

³³Dvs. y -aksen vår er sammenfallende med den stiplete linjen.

³⁴Du skal ikke regne ut verdiene i lign.(66). Bare ta dem for gitt.

Man kan også regne ut den empiriske variansen til x , dvs. S_x^2 , og den empiriske variansen til y , dvs. s_y^2 . Resultatene er:

$$s_x^2 = 188 \quad , \quad s_y^2 = 4\,808\,757.10 \quad (67)$$

Den empiriske **kovariansen** mellom x og y er: ³⁵

$$s_{xy} = 29\,354.24 \quad (68)$$

a) **2.5 %** Regn ut korrelasjonskoeffisienten r_{xy} for observasjonene via de oppgitte størrelsene i oppgaven.

b) **2.5 %** Tolk svaret du fikk for r_{xy} i oppgave **4a**. ³⁶

³⁵Størrelsene i lign.(67) og lign.(68) skal du ikke regne ut. Bare ta dem for gitt.

³⁶Hva sier den numeriske (tallmessige) verdien av r_{xy} fra oppgave **4a** om graden av korrelasjon mellom observasjonene x og y ?

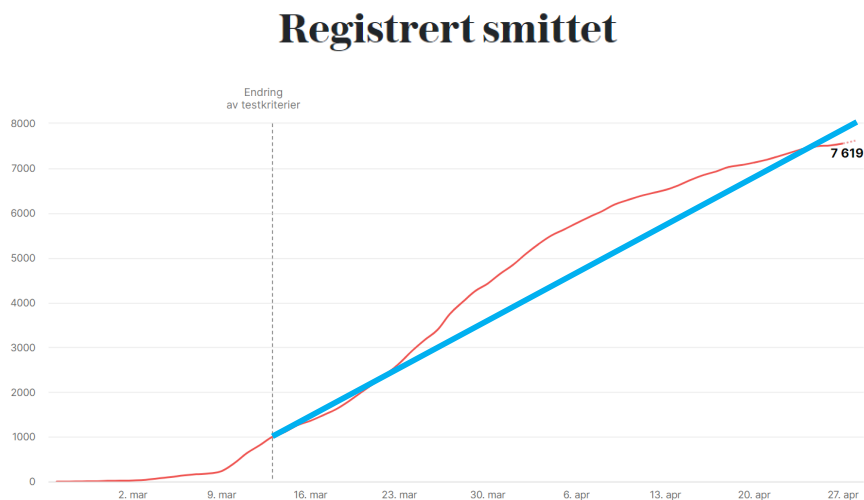
Pga. resultatene fra oppgave 4a og 4b konkluderer vi med at det kan være hensiktsmessig å finne en eksplisitt lineær sammenheng mellom x og y . I den sammenheng skal vi bruke lineær regresjonsanalyse.

- c) 7.5 % Den blå rette linjen i figur 13 viser minste kvadraters regresjonslinje for observasjonene x_i og y_i .

Finn ved regning et analytisk uttrykk for denne lineære **regresjonslinjen**. ³⁷

- d) 5 % Anta nå at den lineære trenden beskrevet av regresjonslinjen fra oppgave 4c holder seg helt frem til i dag, mandag 18. mai 2020.

Ifølge regresjonslinjen, hvor mange smittede er det i Norge i dag, dvs. per 18. mai 2020? ³⁸



Figur 13: Regresjonslinje (blå).

³⁷Dvs. finn formelen for linjen via setningen for minste kvadraters lineære regresjonslinje. Se gjerne kompendiet.

³⁸Mandag 18. mai 2020 tilsvarer dag nr. $x = 69$.

- e) 5% Ifølge regresjonslinjen fra oppgave 4c, når er det 5 000 personer som er smittet av koronaviuset i Norge? ³⁹

Størrelsen *see*, “*sum squared error*”, og størrelsen *sst*, “*sum squared total*”, er i vårt tilfelle gitt ved:

$$sse = 10\,368\,356 \quad , \quad sst = 221\,202\,826 \quad (69)$$

Tallene i lign.(69) skal du ikke regne ut. Bare ta dem for gitt.

- f) 2.5% Bruk lign.(69) og regn ut forklaringskraften r^2 . ⁴⁰

Kommenter svaret. ⁴¹

Stemmer ditt svar med Excel-utskriften i figur 14 nedenfor?

120	SUMMARY OUTPUT					
121						
122	<i>Regression Statistics</i>					
123	Multiple R	0,9762825				
124	R Square	0,9531274				
125	Adjusted R Square	0,9520858				
126	Standard Error	480,00801				
127	Observations	47				
128						
129	ANOVA					
130		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
131	Regression	1	210834480,5	210834480,5	915,0497	1,50857E-31
132	Residual	45	10368346,09	230407,6909		
133	Total	46	221202826,6			
134						
135		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>
136	Intercept	1368,1303	137,8273513	9,926406522	6,55E-13	1090,531784
137	X Variable 1	156,13957	5,161674872	30,24978785	1,51E-31	145,743423

Figur 14: Utskrift fra Excel.

³⁹Du trenger ikke oppgi dato, kun x -verdi.

⁴⁰Bruk 4 desimalers nøyaktighet.

⁴¹Hva sier svaret om hvor stor grad variansen til y kan forklares med x ?