

LØSNING: Eksamen 18. mai 2020

“MAT110 Statistikk 1”, vår 2020

Versjon 01

Oppgave 1: (sannsynlighetsteori - koronatest)

a) 2.5 % Utfallsrommet Ω har 4 mulige utfall: ¹

$P(t, s)$ = sannsynligheten for at
test sier **positiv** og personen er **smittet** (korrekt test) (1)

$P(\bar{t}, s)$ = sannsynligheten for at
test sier **negativ** og personen er **smittet** (falsk negativ test) (2)

$P(t, \bar{s})$ = sannsynligheten for at
test sier **positiv** og personen er **ikke smittet** (falsk positiv test) (3)

$P(\bar{t}, \bar{s})$ = sannsynligheten for at
test sier **negativ** og personen er **ikke smittet** (korrekt test) (4)

¹Den formelt korrekte notasjonen er $P(\{(\bar{t}, s)\})$, men vi forenkler notasjonen til kun $P(\bar{t}, s)$.

b) 5% Komplementsetningen:

$$P(\overline{T} | S) = 1 - \overbrace{P(T | S)}^{= 1} = 1 - 1 = 0 \quad (5)$$

Multiplikasjonssetningen:

$$P(\overline{t}, s) = \overbrace{P(\overline{T} | S)}^{= 0} \overbrace{P(S)}^{= 0.01} = 0 \cdot 0.01 = 0 \quad (6)$$

c) 2.5% Lign.(6) sier at falske negative tester ikke inntreffer (umulig).

d) 5% Oppsplitting av utfallsrom:

$$\begin{aligned} P(T) &= P(T | \overline{S})P(\overline{S}) + P(T | S)P(S) \\ &= \underbrace{0.01}_{\text{oppgitt}} \cdot (1 - \underbrace{0.01}_{\text{oppgitt}}) + \underbrace{1 \cdot 0.01}_{\text{oppgitt}} = 0.0199 \quad , \quad \text{q.e.d.} \end{aligned} \quad (7)$$

e) 5% Bayes lov:

$$P(\overline{S} | T) = \frac{P(T | \overline{S})P(\overline{S})}{P(T)} \quad (8)$$

$$= \frac{0.01 \cdot (1 - 0.01)}{0.0199} = 0.4975 \quad , \quad \text{q.e.d.} \quad (9)$$

f) 5% Siden $P(\overline{S} | T) = 0.4975$ så betyr det at nesten halvparten av testene som utføres på friske personer gir feil resultat.

Det er derfor ikke lurt å teste på friske personer.

■

Oppgave 2: (stokastiske variabler og forventningverdi - aldersfordeling)

a) 7.5 % Forventet levealder på en person som tester positivt på korona:

$$E[X] = \sum_{i=1}^{10} x_i P(X = x_i) \quad (10)$$

$$= \sum_{i=1}^{10} x_i p_i \quad (11)$$

$$\begin{aligned} &= 4.5 \cdot \frac{77}{6791} + 14.5 \cdot \frac{295}{6791} + 24.5 \cdot \frac{1002}{6791} + 34.5 \cdot \frac{1100}{6791} + 44.5 \cdot \frac{1244}{6791} + \\ &+ 54.5 \cdot \frac{1331}{6791} + 64.5 \cdot \frac{789}{6791} + 74.5 \cdot \frac{530}{6791} + 84.5 \cdot \frac{323}{6791} + 94.5 \cdot \frac{102}{6791} \end{aligned}$$

$$= 47.45 \text{ år} \quad , \quad \text{q.e.d.} \quad (12)$$

Varianssentingen: (oppgitt i år²)

$$Var[X] = E[X^2] - E[X]^2 \quad (13)$$

$$= 2624.66 - 47.45^2 = 373.16 \quad , \quad \text{q.e.d.} \quad (14)$$

med tilhørende standardavvik (oppgitt i år)

$$\sigma[X] = \sqrt{Var[X]} = \sqrt{373.16} = 19.32 \quad (15)$$

b) 2.5 % Siden

$$Y \sim N \left[E[Y], \sigma[Y] \right] \quad (16)$$

hvor $E[Y] = E[X] = 47.45$ og $\sigma[Y] = \sigma[X] = 19.32$, så er:

$$\underline{P(Y < 20)} \stackrel{\text{standardiser}}{=} P \left(\frac{Y - E[X]}{\sigma[X]} < \frac{20 - E[X]}{\sigma[X]} \right) \quad (17)$$

$$= P \left(\underbrace{\frac{Y - E[X]}{\sigma[X]}}_{\equiv Z} < \frac{20 - E[X]}{\sigma[X]} \right) \quad (18)$$

$$= P \left(Z < \frac{20 - 47.45}{19.32} \right) \quad (19)$$

$$= P(Z < -1.42) \quad \text{minustegn} \quad (20)$$

$$= 1 - P(Z < 1.42) \quad (21)$$

Tabeloppslag:

$$P(Y < 20) = 1 - 0.9222 = 0.0778 \quad (22)$$

Tolkning:

$$P(X < 20) = \text{sannsynligheten for at en tilfeldig valgt person som} \\ \text{har testet positivt på korona i eksperiment } E' \quad (23)$$

$$\text{er under 20 år, er 0.0778, dvs ca. 7.78\% tester positivt.} \quad (24)$$

- c) 5% La n_i antall smittende i aldergruppe nr. $i = 1, \dots, 10$.
 “Korrigerte” forventningen $E'[X]$:

$$E'[X] = \sum_{i=1}^{10} x_i P(X = x_i) \quad (25)$$

$$= 4.5 \frac{n_1}{N+m} + 14.5 \frac{n_2 + m}{N+m} + 24.5 \frac{n_3}{N+m} + \dots + 94.5 \frac{n_{10}}{N+m} \quad (26)$$

$$= \left(4.5n_1 + 14.5n_2 + 14.5m + 24.5n_3 + \dots + 94.5n_{10} \right) \frac{1}{N+m} \quad (27)$$

$$= \left(4.5n_1 + 14.5n_2 + 24.5n_3 + \dots + 94.5n_{10} \right) \frac{1}{N+m} + \frac{14.5m}{N+m} \quad (28)$$

$$= \underbrace{\frac{4.5n_1 + 14.5n_2 + 24.5n_3 + \dots + 94.5n_{10}}{N}}_{= E[X]} \frac{N}{N+m} + \frac{14.5m}{N+m} \quad (29)$$

$$= E[X] \frac{N}{N+m} + \frac{14.5m}{N+m} \quad (30)$$

$$= \frac{E[X] N + 14.5m}{N+m}, \quad \text{q.e.d.} \quad (31)$$

- d) 5% Bruker lign.(31) og løser med hensyn på m alene:

$$E'[X] = \frac{E[X] N + 14.5m}{N+m} \quad \Big| \cdot N+m \quad (32)$$

$$\Downarrow$$

$$(N+m)E'[X] = E[X] N + 14.5m \quad (33)$$

$$\Downarrow$$

$$NE'[X] + mE'[X] = E[X] N + 14.5m \quad (34)$$

$$\Downarrow$$

$$mE'[X] - 14.5m = E[X] N - NE'[X] \quad (35)$$

$$\Downarrow$$

$$m(E'[X] - 14.5) = N(E[X] - E'[X]) \quad \Big| \cdot \frac{1}{E'[X] - 14.5} \quad (36)$$

$$\Downarrow$$

$$m = \frac{N(E[X] - E'[X])}{E'[X] - 14.5} \quad (37)$$

Sett til slutt inn tall: ($N = 6791$, $\overbrace{E[X] = 47.45}^{\text{oppg. 2a}}$ og $E'[X] = 46$)

$$m = \frac{N(E[X] - E'[X])}{E'[X] - 14.5} \quad (38)$$

$$= \frac{6791(47.45 - 46)}{46 - 14.5} = 313 \quad (39)$$

Altså, mørketallet er $m = 313$ dersom $E'[X] = 46$ år.

e) 5% Ved regning:

Denne oppgaven løses på samme måte som oppgave 2b, men med $E'[X] = 46$ og $\sigma'[X] = 20.07$.²

Siden

$$\frac{20 - 46}{20.07} = -1.30 \quad (40)$$

så får man ved tabelloppslag:

$$P(Y < 20) = 1 - 0.9032 = 0.0968 \quad (41)$$

som er større enn 0.0778 fra oppgave 2a.

Med ord:³

Siden det nå er flere personer i aldersgruppen 10-19 år sammenlignet med oppgave 2b, men ellers likt, så er sannsynligheten i lign.(41) større enn i lign.(22).

■

²Man trenger ikke å repetere alle utregningene fra oppgave 2b.

³Man kan enten vise det ved regning eller argumentere med ord. Man trenger ikke begge deler. Her i løsningsforslaget er det tatt med begge måtene for kompletthets skyld.

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

a) 2.5 % Med $B_0 = B_0(2, 1, 0)$ så er: ⁴

$$I_1 \mid B_0 \sim \text{Bin}[2, p^1 = p] \quad (42)$$

som gir

$$P(I_1 = 1 \mid B_0) = \binom{2}{1} p^1 (1-p)^{2-1} = 2p(1-p) = 2p - 2p^2, \quad \text{q.e.d.} \quad (43)$$

b) 2.5 % Med $B_0 = B_0(2, 1, 0)$ så er:

$$I_1 \mid B_0 \sim \text{Bin}[2, p^1 = p] \quad (44)$$

som gir

$$P(I_1 = 2 \mid B_0) = \binom{2}{2} p^2 (1-p)^{2-2} = 1 \cdot p^2 \cdot 1 = p^2, \quad \text{q.e.d.} \quad (45)$$

c) 2.5 % Med de oppgitte startverdiene $B_0 = B_0(N-1, 1, 0)$ så er:

$$I_1 \mid B_0(N-1, 1, 0) \sim \text{Bin}[N-1, p^1 = p] \quad (46)$$

Dermed er forventningen av antall smittede etter dag 1 gitt ved:

$$E[I_1 \mid B_0] = (N-1)p \quad (47)$$

⁴Husk at: $N-1 = 3-1=2$.

d) 2.5 % Matematisk modell: (oppgitt i oppgaven)

$$R_0 = \beta cD \quad (48)$$

Bruker i tillegg lign.(47), dvs. $R_0 = (N - 1)p$:

$$(N - 1)p = \beta cD \quad (49)$$

Løser med hensyn på p alene og setter inn tall til slutt:

$$p = \frac{\beta cD}{N - 1} \quad (50)$$

$$= \frac{0.5 \cdot 3 \cdot 1}{3 - 1} = 0.75 \quad (51)$$

- e) 10% Sannsynligheten for at hele populasjonen blir smittet, dvs. $Z = N = 3$, er lik 0.01:

$$P(Z = 3 \mid B_0) = 0.01 \quad (52)$$

Med startverdiene $B_0 = B_0(2, 1, 0)$, $N = 3$ personer samt antagelsene i oppgaven så kan vi se på de forskjellige mulige forløpene. Da finner vi at det forløpet som vist i figuren i oppgaveteksten er det som gir max n , dvs. max levelengde på epedemien. Epedemien dør da ut i uke 3, dvs. $I_3 = 0$, og dermed $I_3 = I_4 = I_5 = \dots = 0$.⁵

$$Z = I_0 + I_1 + I_2 + \overbrace{I_3 + I_4 + \dots}^{=0} = I_0 + I_1 + I_2 \quad (53)$$

Med $B_0 = B(2, 1, 0) = \{S_0 = 2, I_0 = 1, R_0 = 0\}$ og $Z = 3$:

$$3 = 1 + I_1 + I_2 \quad (54)$$

$$\Downarrow \quad (55)$$

$$I_1 + I_2 = 2 \quad (56)$$

Innsatt i lign.(52):

$$P(I_1 + I_2 = 2 \mid B_0) = 0.01 \quad (57)$$

To kombinasjoner hvor $I_1 + I_2 = 2$:⁶

$$P\left(I_1 = 1, I_2 = 1 \underbrace{\vee}_{\text{eller}} I_1 = 2, I_2 = 0 \mid B_0\right) = 0.01 \quad (58)$$

⁵Her er et eksempel å et mulig forløp hvor epedemien dør ut tidligere:

Dersom begge $S_0 = 2$ hadde blitt smittet i uke 1, så er det ikke flere som kan smittes. Begge hadde blitt immune påfølgende uke, og epedemien hadde dødd ut i uke $n = 2$, og $I_2 = I_3 = I_4 = \dots = 0$.

⁶Kombinasjonen $\{I_1 = 0, I_2 = 2\}$ er umulig med Markov tilnærmelsen og antagelse 4 i oppgaven. Med $I_1 = 0$ så stopper smitten i uke 1, og man kan ikke få smittede personer i uke 2 eller senere.

Addisjonssetningen: ⁷

$$P(I_1 = 1, I_2 = 1 \mid B_0) + P(I_1 = 2, I_2 = 0 \mid B_0) = 0.01 \quad (59)$$

Definisjonen av betinget sannsynlighet: (med $S_0 = 2$ og $I_0 = 1$)

$$\begin{aligned} & P(I_2 = 1 \mid I_1 = 1, B_0) \overbrace{P(I_1 = 1 \mid B_0)}^{\sim \text{Bin}[S_0=2, p^1]} \\ & + P(I_2 = 0 \mid I_1 = 2, B_0) \underbrace{P(I_1 = 2 \mid B_0)}_{\sim \text{Bin}[S_0=2, p^1]} = 0.01 \end{aligned} \quad (60)$$

Vi bruker at sannsynlighetene i lign.(60) er binomialfordelte med smitte-sannsynlighet p :

$$P(I_2 = 1 \mid I_1 = 1, B_0) \overbrace{\binom{2}{1} p^1 (1-p)^{2-1}}^{= 2p-2p^2, \text{ oppg. 3a}} \quad (61)$$

$$+ P(I_2 = 0 \mid I_1 = 2, B_0) \underbrace{\binom{2}{2} p^2 (1-p)^{2-2}}_{= p^2, \text{ oppg. 3b}} = 0.01 \quad (62)$$

Forenkler:

$$\underbrace{P(I_2 = 1 \mid I_1 = 1, B_0)}_{= p} (2p - 2p^2) + \underbrace{P(I_2 = 0 \mid I_1 = 2, B_0)}_{= 1} p^2 = 0.01 \quad (63)$$

⁷Begivenhetene $\{I_1 = 1, I_2 = 1\}$ og $\{I_1 = 2, I_2 = 0\}$ er disjunkte.

i) Regner ut $P(I_2 = 1 \mid I_1 = 1, B_0)$: (den første sannsynligheten i lign.(63))

Med $B_0 = B_0(2, 1, 0)$ og $I_1 = 1$:

$$\left. \begin{aligned} S_1 &= S_0 - I_1 = 2 - 1 = 1 \\ I_1 &= 1 \\ R_1 &= R_0 + I_0 = 0 + 1 = 1 \end{aligned} \right\} B_1(1, 1, 1) \quad (64)$$

hvor vi har brukt de oppgitte ligningene i oppgaven. ⁸

Sannsynligheten $P(I_2 = 1 \mid I_1 = 1, B_0)$ i lign.(63) kan dermed skrives:

$$P(I_2 = 1 \mid I_1 = 1, B_0) = P(I_2 = 1 \mid B_1, \cancel{B_0}) \quad (65)$$

$$= P(\underbrace{I_2 = 1}_{\sim \text{Bin}[s_1=1, p^2]} \mid B_1) \quad (66)$$

$$= \binom{1}{1} p^1 (1-p)^{1-1} \quad (67)$$

$$= p \quad (68)$$

hvor vi kan fjerne $\cancel{B_0}$ i betingelsen pga. Markov tilnærmelsen, dvs. prosessen er kun avhengig av hva som skjedde bare én uke før, altså “korttidshukommelse”.

I lign.(66) har vi brukt binomialfordelingen med $B_1 = B_1(1, 1, 1)$. ⁹

⁸Man kan også innse lign.(64) ved hjelp av figuren tilhørende oppgave 3 og antagelsene i oppgaven.

⁹Ved hjelp av figuren tilhørende oppgave 3 og antagelsene i oppgaven kan man også argumentere for at $P(I_2 = 1 \mid I_1 = 1, B_0) = p$, men vi velger den matematiske tilnærmelsen her.

ii) Regner ut $P(I_2 = 0 \mid I_1 = 2, B_0)$: (den andre sannsynligheten i lign.(63))

Med $B_0 = B_0(2, 1, 0)$ og $I_1 = 2$:

$$\left. \begin{aligned} S_1 &= S_0 - I_1 = 2 - 2 = 0 \\ I_1 &= 2 \\ R_1 &= R_0 + I_0 = 0 + 1 = 1 \end{aligned} \right\} B_1(0, 2, 1) \quad (69)$$

hvor vi har brukt de oppgitte ligningene i oppgaven. ¹⁰

Sannsynligheten $P(I_2 = 0 \mid I_1 = 2, B_0)$ i lign.(63) kan dermed skrives:

$$P(I_2 = 0 \mid I_1 = 2, B_0) = P(I_2 = 0 \mid B_1, \cancel{B_0}) \quad (70)$$

$$= P(\underbrace{I_2 = 0}_{\sim \text{Bin}[s_1=0, p^2]} \mid B_1) \quad (71)$$

$$= \binom{0}{0} (p^2)^0 (1 - p^2)^{0-0} \quad (72)$$

$$= 1 \quad (73)$$

hvor vi kan fjerne $\cancel{B_0}$ i betingelsen pga. Markov tilnærmelsen, dvs. prosessen er kun avhengig av hva som skjedde bare én uke før, altså “korttidshukommelse”.

I lign.(71) har vi brukt binomialfordelingen med $B_1 = B_1(0, 2, 1)$. ¹¹

¹⁰Man kan også innse lign.(69) ved hjelp av figuren tilhørende oppgave 3 og antagelsene i oppgaven.

¹¹Ved hjelp av figuren tilhørende oppgave 3 og antagelsene i oppgaven kan man også argumentere for at $P(I_2 = 0 \mid I_2 = 2, B_0) = 1$, men vi velger den matematiske tilnærmelsen her.

Setter inn lign.(68) og lign.(73) inn i lign.(63):

$$p \cdot (2p - 2p^2) + 1 \cdot p^2 = 0.01 \quad (74)$$

$$\Downarrow$$

$$2p^2 - 2p^3 + p^2 = 0.01 \quad (75)$$

$$\Downarrow$$

$$3p^2 - 2p^3 = 0.01 \quad (76)$$

Fra fotnoten:

$$p = 1.4977 \quad , \quad p = -0.0567 \quad , \quad \underline{p = 0.0589} \quad (77)$$

som gir at $p = 0.0589$, siden sannsynligheten må ligge mellom 0 og 1.

Fra lign.(49) i oppgave **3d**:

$$R_0 = (N - 1)p \quad (78)$$

$$\Downarrow$$

$$\beta c' D = (N - 1)p \quad \Big| \cdot \frac{1}{\beta D} \quad (79)$$

$$\Downarrow$$

$$c' = \frac{(N - 1)p}{\beta D} \quad (80)$$

Tall: $(p = 0.0589, \beta = 0.5, D = 1 \text{ og } N = 3)$

$$c' = \frac{(N - 1)p}{\beta D} = \frac{(3 - 1)0.0589}{0.5 \cdot 1} = 0.2356 \quad (81)$$

Isolasjonen må være slik at kontakthypigheten blir redusert til $c' = 0.2356$, dvs. man treffer i gjennomsnitt 0.23556 personer i uken, som tilsvarer ca. 1 personer per måned.

■

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

- a) 2.5 % Uttrykket for korrelasjonskoeffisienten finner vi i kompendiet. Det gir:

$$\underline{\underline{r_{xy}}} = \frac{s_{xy}}{s_x \cdot s_y} \quad (82)$$

$$= \frac{29\,354.24}{\sqrt{188} \cdot \sqrt{4\,808\,757.10}} = \underline{\underline{0.9763}} \quad (83)$$

- b) 2.5 % At $r_{xy} = 0.9763$ betyr at det er en sterk positiv korrelasjon mellom x og y , dvs. store x hører sammen med store y og omvendt. Det er altså en sterk lineær sammenheng mellom x og y med positivt stigningstall.

- c) 7.5 % Vi skal finne regresjonslinjen. Formelen for minste kvadraters lineære regresjonslinje finner man i kompendiet:

$$\hat{y} = \hat{\alpha} + \hat{\beta}x \quad (84)$$

hvor parametrene $\hat{\beta}$ og $\hat{\alpha}$ er: (dropper benevningen)

$$\underline{\hat{\beta}} = \frac{s_{xy}}{s_x^2} = \frac{29\,354.24}{188} = \underline{156.14} \quad (85)$$

$$\underline{\hat{\alpha}} = \bar{y} - \hat{\beta}\bar{x} = 4959.34 - 156.14 \cdot 23 = \underline{1368.13} \quad (86)$$

Minste kvadraters lineære regresjonslinje $\hat{y} = \hat{\alpha} + \hat{\beta}x$ blir dermed:

$$\underline{\underline{\hat{y} = 1368.13 + 156.14x}} \quad (87)$$

- d) 2.5 % Datoen 18. mai 2020 tilsvarer $x = 69$.
Bruker regresjonslinjen $\hat{y} = \hat{y}(x)$ fra oppgave 4c, dvs. lign.(87):

$$\underline{\hat{y}(69)} = \left(1368.13 + 156.14 \cdot 69 \right) = \underline{12\,141.79} \quad (88)$$

Ifølge regresjonslinjen kommer det til å være 12 142 personer som er smittet i dag, 18. mai 2020.

- e) 5 % Ifølge regresjonslinjen i lign.(87), når $\hat{y}(x) = 5\,000$, er da:

$$5\,000 = \hat{\alpha} + \hat{\beta}x \quad (89)$$

Løser med hensyn på x : ($\alpha = 1368.13$ and $\hat{\beta} = 156.14$)

$$\underline{x} = \frac{5\,000 - \hat{\alpha}}{\hat{\beta}} \quad (90)$$

$$= \frac{5\,000 - 1368.13}{156.14} = \underline{23.26} \quad (91)$$

Ifølge regresjonslinjen er det 5 000 personer i Norge som er smittet av korona dag nr. 23. ¹²

¹²Man behøver ikke å skrive at dag nr. 23 tilsvarer 5. april 2020.

f) 2.5% Bruker definisjonen av forklaringskraften r^2 :

$$r^2 = 1 - \frac{sse}{sst} \quad (92)$$

$$= 1 - \frac{10\,368\,356}{221\,202\,826} = 0.9531 \quad (93)$$

som stemmer med Excel-utskriften.

Kommentar:

At $r^2 = 0.9531$ betyr at 95.31 % av variasjonen til y , antall smittede, kan forklares av x , dvs. tiden.

■