NTNU

Kunnskap for en bedre verden

Institutt for matematiske fag

Eksamensoppgave i
TMA4245 Statistikk. LØSNINGSFORSLAG

Faglig kontakt under eksamen:

Tlf:

Eksamensdato: 15. mai 2020

Eksamenstid (fra–til): 09:00–13:00

Hjelpemiddelkode/Tillatte hjelpemidler: A: Alle hjelpemidler tillatt.

Målform/språk: bokmål

Antall sider: 3

Antall sider vedlegg: 0

Kontrollert av:

Informasjon om trykking av eksamensoppgave	
Originalen er:	
1-sidig ☐	**2-sidig** ☒
sort/hvit ☒	**farger** ☐
skal ha flervalgskjema ☒	

Dato Sign

Oppgave 1 Flervalgsoppgaver, hendelser og sannsynlighet.

a) Det skraverte området i venndiagrammet er $C \cap A'$

b) $\mathrm{P}(B|A) = \frac{\mathrm{P}(B \cap A)}{\mathrm{P}(A)} = \frac{20/100}{55/100} = 0.364$

c) Det er flere mulige fremgangsmåter for å løse en slik oppgave. For eksempel kan du regne ut P(blå i første trekning, blå i andre, rød i tredje) + P(blå i første trekning, rød i andre, blå i tredje) + P(rød i første trekning, blå i andre, blå i tredje), som ved trekning *uten* tilbakelegging tilsvarer

$$\frac{14}{20} \cdot \frac{13}{19} \cdot \frac{6}{18} + \frac{14}{20} \cdot \frac{6}{19} \cdot \frac{13}{18} + \frac{6}{20} \cdot \frac{14}{19} \cdot \frac{13}{18} = 3 \cdot \frac{14 \cdot 13 \cdot 6}{20 \cdot 19 \cdot 18} = 0.479$$

d) Som i forrige deloppgave kan du for eksempel regne ut P(blå i første trekning, blå i andre, rød i tredje) + P(blå i første trekning, rød i andre, blå i tredje) + P(rød i første trekning, blå i andre, blå i tredje), som ved trekning *med* tilbakelegging tilsvarer

$$\frac{14}{20} \cdot \frac{14}{20} \cdot \frac{6}{20} + \frac{14}{20} \cdot \frac{6}{12} \cdot \frac{14}{20} + \frac{6}{20} \cdot \frac{14}{20} \cdot \frac{14}{20} = 3 \cdot \frac{14 \cdot 14 \cdot 6}{20^3} = 0.441$$

Oppgave 2

a) $\mathrm{E}[Y] = \frac{1}{3}(2 \cdot 1 - 5 \cdot 1) = -1$

b) $\mathrm{Var}[Y] = \frac{1}{3^2}(2^2 \cdot 2 + 5^2 \cdot 3 - 2 \cdot 2 \cdot 5 \cdot 1) = 7$

Oppgave 3 Poisson-prossess

a) Nei. Antall kunder som ankommer mellom klokka 9.00 og 10.00 er Poisson-fordelt med forventningsverdi $\lambda = 2$, og antall kunder som ankommer mellom klokka 12:00 og 13:00 er Poissonfordelt med forventningsverdi $\lambda = 2$. Dermed er det ingen grunn til å forvente forskjellig antall kunder i tidsperiodene 9.00 - 10.00 og 12:00 - 13:00.

b) Vi lar X betegne antall ankomster mellom klokka 10:00 og 12:00. X er da Poisson-fordelt med parameter $2 \cdot \lambda = 4$, uavhengig av hva som har skjedd før klokka 10:00. Dermed finner vi

$$\begin{aligned}\mathrm{P}(X \geq 4) &= 1 - \mathrm{P}(X \leq 3) \\ &= 1 - (\mathrm{P}(X = 0) + \mathrm{P}(X = 1) + \mathrm{P}(X = 2) + \mathrm{P}(X = 3)) \\ &= 0.567\end{aligned}$$

Oppgave 4

a) $\mathrm{P}(3 < X \leq 5) = \mathrm{P}(X \leq 5) - \mathrm{P}(X \leq 3) = F(5) - F(3) = 1 - (3/5)^2 - (1-1) = 1 - (3/5)^2 = 0.640$

b) $\mathrm{P}(X > 6 | X > 4) = \frac{\mathrm{P}(X>6 \cap X>4)}{\mathrm{P}(X>4)} = \frac{\mathrm{P}(X>6)}{\mathrm{P}(X>4)} = \frac{1-F(6)}{1-F(4)} = \frac{(3/6)^2}{(3/4)^2} = 0.444.$

Oppgave 5 Meslingvaksine

a) Hver X_i kan antas Bernoulli(p)-fordelt (eller tilsvarende Binomisk(1,p)-fordelt), siden vi trekker et relativt lite antall personer ($n = 100$, uten tilbakelegging) fra en stor populasjon. Parameteren p representerer andelen vaksinerte 16-åringer i Norge i 2020.

b) Vi ønsker å kunne konkludere med at $p < 0.91$. Derfor er alternativhypotesen $H_1 : p < 0.91$. Da må nullhypotesen være $H_0 : p = 0.91$ (evt $H_0 : p \geq 0.91$).

c) Fra sentralgrenseteoeremet følger det at $\hat{P}$ er tilnærmet normalfordelt, og dermed tilnærmet $N(p_0, p_0(1-p_0)/100)$-fordelt dersom nullhypotesen er sann. Innsatt $p_0 = 0.91$ får vi en $N(0.91, 0.000819)$-fordeling, der 0.000819 er variansen. (Alternativt kan man oppgi standardavviket som er 0.0286.)

d) Vi bruker en testobservator Z som er (tilnærmet) standard normalfordelt dersom nullhypotesen er sann. Testobservatoren er gitt ved $Z = (\hat{P} - 0.91)/0.0286$, evt $Z = (\hat{P} - p_0)/\sqrt{p_0(1-p_0)/100}$, der $p_0 = 0.91$. Dette er en ensidig test og vi må forkaste hvis vi observerer $z < -z_{0.05} = -1.645$. Vi regner ut $z = -0.699 > -z_{0.05} = -1.645$. Konklusjonen er at vi ikke kan forkaste H_0, og vi har ikke grunn til å tro at andelen vaksinerte 16-åringer i Norge har blitt lavere.

Oppgave 6 Lineær regresjon

a) Det virker rimelig å anta en lineær sammenheng mellom BMI og fettprosent. Det virker også også rimelig å anta at variasjonen i fettprosent (varians σ^2) er den samme for alle verdier av BMI. Videre er det ikke noe mønster i avvikene fra regresjonslinjen og det er rimelig å anta uavhengige residualer. (Det er vanskelig å vurdere antagelsen om normalfordeling fra figuren, men den er ingen indikasjoner på at normalfordelingen er urimelig (slik som outliers).)

b) For BMI = 20 er predikert fettprosent 11.1, og for BMI = 30 er predikert fettprosent 31.3. Ingen av personene i studien har BMI på 30 eller høyere, mens fem av personene i studien har BMI nær 20. Derfor er vi mer usikre på prediksjonen for personer med BMI = 30. Det er også mulig å bruke et mer formelt argument basert på uttrykket for bredden av konfidensintervallene eller prediksjonsintervallene.

Oppgave 7 To utvalg

a) (Flervalg) Forventningsverdien til den første estimatoren er $\mu_1 - \mu_2$, forventnigsverdien til den andre estimatoren er $\mu_1 - \mu_2$, men forventningsverdien til den tredje estimatoren er $\mu_1 - \mu_2 - \mu_2 = \mu_1 - 2\mu_2$. Derfor er det bare $\hat{d}$ og d^* som er forventningsrette.

b) Den første estimatoren er forventningsrett med varians $\frac{1}{10^2} \cdot 10 \cdot 4 + \frac{1}{25^2} \cdot 25 \cdot 4 = 0.56$, mens den andre estimatoren er forventningsrett med varians $\frac{1}{10^2} \cdot 10 \cdot (4+4) = 0.8$. Den tredje estimatoren er ikke forventningsrett. Vi foretrekker en forventningsrett estimator med lav varians, derfor foretrekker vi den første estimatoren $\hat{d}$.

c) Konfidensintervallet er $[0.449, 3.382]$. Siden variansen er kjent i begge populasjoner kan vi ta utgangspunkt i en standard normalfordelt stokastisk variabel $Z = (\hat{d} - d)/\sqrt{0.56}$. Estimatoren gir estimatet $\hat{d} = 1.9152$ og estimert standardfeil $\sqrt{0.56} = 0.748$. Ved å løse $P(-z_{0.025} \leq Z \leq z_{0.025}) = 0.95$ finner vi et tosidig intervall der nedre og øvre grenser for konfidensintervallet er $z_{0.025} = 1.96$ standardfeil unna estimatet. Det vil si at et 95% konfidensintervall er gitt ved $[\hat{d} - 1.96 \cdot 0.748, \hat{d} + 1.96 \cdot 0.748] = 0.95$, der $\hat{d} = 1.9152$.

d) Siden variansene er de samme i begge populasjoner bør vi gjøre 15 nye observasjoner fra Populasjon 1, og ingen nye observasjoner fra Populasjon 2, slik at vi får like mange observasjoner fra hver populasjon.

Bredden av intervallet er proposjonal med

$$1/(\text{antall obs. fra Pop. 1}) + 1/(\text{antall obs. fra Pop. 2}).$$

Dersom vi lar n_1 være antall nye observasjoner fra Populasjon 1, og $15 - n_1$ antall nye observasjoner fra Populasjon 2, så vil $n_1 = 15$ minimere funksjonen $1/(10+n_1) + 1/(25+(15-n_1))$ og dermed gi det smaleste konfidensintervallet.