Universitetet i Stavanger

EKSAMEN I: STA100 SANNSYNLIGHETSREGNING OG STATISTIKK

VARIGHET: 5 TIMER (inkludert tid til å klargjøre og levere besvarelsen).
TID: 9:00-14:00 den 14. MAI 2020.
EKSAMEN BESTÅR AV 3 OPPGAVER PÅ 5 SIDER.

HJELPEMIDLER: Alle tekniske hjelpemidler er lovlige. Det er *ikke* lov å få hjelp av andre personer i arbeidet med eksamensoppgaven.

KONTAKT UNDER EKSAMEN:
Spørsmål om tolkning av eksamensoppgavene: 51 83 22 55 / 48 44 90 14 / 51 83 18 74
Administrativ støtte: 51 83 17 15 / 51 83 31 33 / 92 81 65 97 / 91 78 67 16
Teknisk støtte: 51 83 20 30 / 51 83 20 14

Dersom noe går galt i Inspera slik at du ikke får levert oppgaven din, kan du sende eksamensbesvarelsen din direkte til eksamenTN@uis.no. Husk å føre på kandidatnummer og emnekode i emnefeltet på e-posten.

Du kan se nedtelling øverst på siden. Her vil du også finne ditt kandidatnummer.
Dersom du har fått innvilget ekstra tid på eksamen vil dette være lagt til din bruker.

MERK ANGÅENDE BESVARELSEN:

Du kan velge å skrive besvarelsen på blanke ark, linjerte ark, ruteark, eller du kan skrive på forhåndslaget svarark med ruter som du finner på Canvas. Du kan også lage besvarelsen i Latex eller annet program om du ønsker det. Besvarelsen skal leveres som en samlet pdf-fil.

Skriv kandidatnummer på hvert ark. Nummerer sidene og skriv totalt antall sider på forsiden. Husk at du ikke må skrive navn eller studentnummer på noen av arkene du skal levere inn. Inspera vil håndtere din identitet og sikre anonym vurdering.

NB! Det gjøres oppmerksom på at du ved semesterregistreringen, har signert digitalt på at du har satt deg inn i regler for plagiering, og regler for fusk som omfattes av eksamensforskriften. På første side i besvarelsen skal du skrive følgende tekst:

Jeg bekrefter å ha fulgt bestemmelsene i eksamensforskriften. Jeg er innforstått med at fusk eller forsøk på fusk vil bli sanksjonert.

Oppgave 1

For å avgjøre om en person er smittet av korona (COVID19) tas en prøve fra slimhinnene i nese/svelg, og denne prøven analyseres på en PCR-maskin. Når PCR-analysen viser at en person har korona sier man at testen er positiv, og når analysen viser at en person ikke har korona sier man at testen er negativ.

I løpet av en uke ble 1311 personer med symptomer som kan tyde på koronasmitte testet i Stavanger-regionen, og 59 av disse testet positivt (dvs testen viste at de hadde korona).

a) Forklar hvilke forutsetninger som må være oppfylte for at antall positive tester blant prøver tatt fra n personer med symptomer er binomisk fordelt.

   Anta at forutsetningene for binomisk fordeling er oppfylte. Ut fra dataene gitt over, finn et estimat og et 95% konfidensintervall for andel smittede blant de med symptomer.

I en senere uke ble 1156 personer med symptomer testet, og 40 av disse testet positivt.

b) Sett opp en krysstabell som gir en oversikt over antall positive og antall negative tester i de to ukene vi har data for (informasjonen gitt rett over og før punkt a)).

   Utfør en kjikvadrattest for dataene i tabellen. Bruk 5% nivå.

   Forklar hva resultatet av testen i praksis betyr.

Et laboratorium har kapasitet til å teste inntil 200 prøver per døgn. La p være andel personer som er positive i befolkningsgruppen som laboratoriet mottar prøver fra, og la nå n betegne antall prøver laboratoriet tester. La videre Y betegne antall positive prøver blant de n prøvene.

c) Under hvilke antagelser kan vi anta at Y er tilnærmet Poisson-fordelt med $\lambda = np$ (og $t = 1$)? (Anta at disse antagelsene er oppfylte i resten av dette punktet.)

   Dersom $n = 200$ og $p = 0.03$, regn ut $P(Y = 4)$ og $P(Y \geq 4)$.

   I løpet av en uke tester laboratoriet 1156 prøver. Hva er sannsynligheten for at minst 40 av disse er positive dersom $p = 0.03$?

I tiden fremover vil omfattende testing for å finne smittede og å kartlegge smittesituasjonen være et viktig ledd i arbeidet med å håndtere pandemi-situasjonen. Et grep man vurderer å ta er å jevnlig teste et stort antall tilfeldig valgte personer fra hele befolkningen og se hvor stor andel som er smittet. Slik kan man følge utviklingen i andelen smittede i hele befolkningen over tid. En mulighet er da å bruke statistisk prosesskontroll som et verktøy for å monitorere utviklingen over tid.

La nå p betegne andelen smittede i hele befolkningen, og anta i punkt d) at denne holder seg uendret.

d) Vis at sannsynligheten er tilnærmet lik α for å få en observert andel, $\hat{p}$, *utenfor* intervallet $[p - z_{\alpha/2}\sqrt{p(1-p)/n}, p + z_{\alpha/2}\sqrt{p(1-p)/n}]$ ved testing av n tilfeldig valgte personer. (Anta at n og p er slik at tilnærming til normalfordeling er ok.)

La Z betegne antall ganger man tester n tilfeldig valgte personer inntil første gang man får en observert andel utenfor intervallet $[p - z_{\alpha/2}\sqrt{p(1-p)/n}, p + z_{\alpha/2}\sqrt{p(1-p)/n}]$. Forklar hvilken fordeling Z har.

Forklar hvorfor ARL $= 1/\alpha$ dersom vi bruker intervallgrensene over som grenser for et $\hat{p}$-diagram.

Strategien for eventuell tilfeldig testing er ikke klar enda, men vi skal se på et tenkt scenario. Anta at man hver uke velger å teste en stikkprøve på 10 000 tilfeldig valgte personer i befolkningen, og at utgangspunktet er en situasjon hvor 0.2% er smittet, dvs $p = 0.002$. Man ønsker med prosesskontroll å avdekke endringer fra dette nivået. Anta videre at utviklingen de 12 første ukene med slik testing blir som angitt i tabellen under. Tabellen angir antall smittede blant de 10 000 som testes hver uke.

uke	1	2	3	4	5	6	7	8	9	10	11	12
antall smittede	17	13	14	18	23	21	26	25	27	29	27	28

e) Regn ut kontrollgrensene for et $\hat{p}$-diagram med ARL $= 50$. Du kan ta utgangspunkt i at $p = 0.002$.

Lag et kontrolldiagram med de beregnede grensene, og plott verdiene i tabellen over inn i diagrammet. Kommenter resultatet.

Forklar hva ARL $= 50$ betyr i praksis. Hva kan være en grunn for å bruke en såpass lav ARL-verdi i denne situasjonen?

Oppgave 2

Rekkevidden til en elbil er antall kilometer bilen kan kjøre fra batteriet er fulladet til det er tomt. Hanna eier en el-bil som har en rekkevidde som er normalfordelt med forventning 245 km og standardavvik 25 km.

Hanna studerer ved UiS, men kommer fra Kristiansand, og kjører derfor av og til strekningen mellom Stavanger og Kristiansand som normalt er 234 km. Imidlertid er det for tiden i lengre perioder omkjøring via Kvinesdal pga veiarbeid, strekningen blir da 252 km. Hanna starter alltid bilturen med fulladet batteri.

a) Hva er sannsynligheten for at Hanna har nok rekkevidde til å kjøre fra Stavanger til Kristiansand når det ikke er omkjøring (dvs når strekningen er 234 km)?

Hva er sannsynligheten for at Hanna *ikke* har nok rekkevidde til å kjøre fra Stavanger til Kristiansand når det er omkjøring (dvs når strekningen er 252 km)?

Hanna planlegger å kjøre strekningen mellom Stavanger og Kristiansand seks ganger i løpet av et semester. Fire av disse turene vil foregå mens det ikke er omkjøring og to av turene med omkjøring. Hva er sannsynligheten for at hun vil oppleve å ha for kort rekkevidde nøyaktig en gang i løpet av disse seks turene?

En forbrukerorganisasjon tester rekkevidde på nye elbiler. For en ny bil på markedet har de gjort ti målinger av rekkevidden under kjørebetingelser som er lignende de betingelsene produsenten oppgir at deres oppgitte rekkevidde gjelder for. Resultatet av forbrukerorganisasjonen sine målinger ble:

295 308 318 298 312 307 293 291 299 321

Det oppgis at $\sum_{i=1}^{10} x_i = 3042$ og $\sum_{i=1}^{10}(x_i - \bar{x})^2 = 1005.6$

Forbrukerorganisasjonen har ofte erfart at rekkevidden de måler er lavere enn den forventede rekkevidden produsenten oppgir. Forbrukerorganisasjonen vil avgjøre med en hypotesetest om de kan konkludere at også denne bilen har lavere forventet rekkevidde enn oppgitt. Produsenten oppgir en forventet rekkevidde på 310 km.

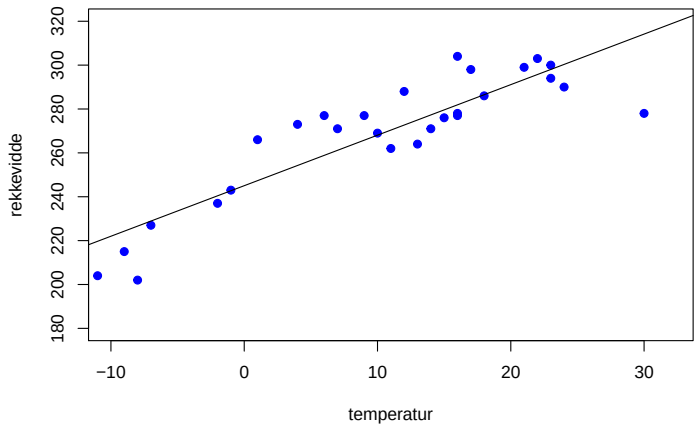
b) Formuler problemstillingen som en hypotesetest.

Forklar hvilke antagelser du må gjøre for å utføre testen.

Utfør testen på 5% nivå.

Forklar hva resultatet av testen betyr i praksis.

Rekkevidden til en elbil avhenger av en rekke faktorer. En faktor som antas å ha betydning er temperatur. Vi skal her se på data fra registreringer av rekkevidde oppnådd med samme elbil ved normal kjøring på dager med ulik temperatur. Et plott av gjennomsnittlig temperatur (i °C) i løpet av kjøreturen og rekkevidde (i antall km) er vist under. En lineær regresjonslinje er også tegnet inn i plottet. Regresjonslinjen tegnet inn i plottet er estimert fra regresjonsmodellen $Y = \alpha + \beta x + e$ der vi antar at $e \sim N(0, \sigma)$ og vi antar at feilleddene $e_1, \ldots, e_n$ for ulike målinger er uavhengige.



Deler av datautskrifter fra R for denne regresjonsmodellen er vist under.

```
Coefficients:
             Estimate   Std. Error   t value    Pr(>|t|)
(Intercept)  245.0341       3.5946     68.168     < 2e-16 ***
temperatur     2.3036       0.2397      9.611    4.81e-10 ***

Residual standard error: 13.76 on 26 degrees of freedom
Multiple R-squared: 0.7804,  Adjusted R-squared: 0.7719
F-statistic: 92.38 on 1 and 26 DF,  p-value: 4.808e-10

> confint(regmod)
                   2.5%       97.5%
(Intercept)     237.645     252.423
temperatur        1.811       2.796
```

c) Gi en praktisk tolkning av α og β i regresjonsmodellen i denne situasjonen.

Skriv ned den estimerte regresjonslinja, og regn ut estimert forventet rekkevidde på en dag med 15 grader.

Gi praktisk tolkning av hva vi tester med hypotesetesten $H_0: \beta = 0$ mot $H_1: \beta \neq 0$ og forklar hva resultatet av testen blir i denne situasjonen.

Bruk informasjon i datautskriften til å regne ut korrelasjonen mellom temperatur og rekkevidde.

I Stavanger er forskjellen i gjennomsnittstemperaturen mellom mai og januar ca 8 grader. Hvor stor er estimert forskjell i forventet rekkevidde mellom en typisk dag i januar og en typisk dag i mai (anta 8 grader forskjell)? Finn også et 95% konfidensintervall for forskjellen i forventet rekkevidde.

Oppgave 3

Hvert år samler Studiebarometeret inn data om studentenes oppfatninger om kvalitet i studieprogrammer ved norske høyskoler og universiteter. En av påstandene som studentene som deltar i undersøkelsen bes om å gi tilbakemelding på er: *"Jeg er, alt i alt, tilfreds med studieprogrammet jeg går på."* Tilbakemeldingen skal gis i form av et tall fra 1 til 5, der 1 betyr "ikke enig" og 5 betyr "helt enig".

La X være tilbakemeldingen på denne påstanden fra en tilfeldig student ved UiS. Dersom vi antar at tallene fra Studiebarometeret for 2019 gir den sanne fordelingen til X er fordelingen:

x	1	2	3	4	5
$P(X = x)$	0.04	0.08	0.21	0.35	0.32

La Y være tilbakemeldingen på den samme påstanden fra en tilfeldig student ved et annet lærested. Anta at fordelingen til Y er:

y	1	2	3	4	5
$P(Y = y)$	0.03	0.07	0.21	0.36	0.33

a) Regn ut $\mathrm{E}(X)$, $\mathrm{Var}(X)$ og $\mathrm{E}(Y)$.

Regn ut sannsynligheten for at en tilfeldig student ved det andre lærestedet er mer fornøyd enn en tilfeldig student ved UiS. Dvs finn $P(Y > X)$.

I realiteten kjenner vi ikke sannsynlighetsfordelingen til X og Y, tallene gitt i tabellen over er kun estimater basert på tall fra de studentene som valgte å svare på undersøkelsen. Dette betyr også at forventningsverdiene $\mu_X = \mathrm{E}(X)$ og $\mu_Y = \mathrm{E}(Y)$ er ukjente.

En oppsummering av tallene fra de to lærestedene for den aktuelle påstanden er at ved UiS så svarte $n_X = 1384$ studenter på påstanden, gjennomsnittsresultatet ble $\bar{x} = 3.83$ og utvalgsstandardavviket ble $s_x = 1.09$. Tilsvarende tall fra det andre lærestedet var $n_y = 987$, $\bar{y} = 3.89$ og $s_y = 1.04$.

b) Utfør en statistisk analyse for å undersøke om det er forskjell i forventet tilfredshet blant studentene ved de to lærestedene (dvs, undersøk om μ_X og μ_Y er forskjellige.) Du må selv avgjøre hva som vil være en hensiktsmessig fremgangsmåte. Forklar hva resultatet av analysen betyr i praksis. Spesifiser hvilke antagelser du gjør og vurder om disse er rimelige.

I vinter var der mange oppslag i media hvor det ble fremstilt som at studentene ved UiS var lite fornøyde basert på gjennomsnittscoren på 3.8 på påstanden omhandlet i denne oppgaven. Fokus i disse oppslagene var at 3.8 var en av de lavest gjennomsnittscorene blant lærestedene i landet. Mange sammenlignbare steder hadde en score på 3.9 eller 4.0. Basert på det du har regnet på i denne oppgaven, er der grunnlag for å si at studenter ved UiS er mindre fornøyde enn et lærested med gjennomsnittlig score på 3.9?