SOSGEO1120 - Eksamen Våren 2021

- Eksamen 14. mai kl.10:00-14:30.
- Eksamensoppgaven er på norsk. Du kan svare på eksamen på norsk, svensk, dansk eller engelsk.

Informasjon om oppgavesettet

Oppgavesettet består av en rekke ulike type oppgaver. I en god del av oppgavene vil det være kun ett spørsmål, mens i andre oppgaver vil det være flere spørsmål du skal besvare. Noen oppgaver krever håndregning før svaret avgis.

OBS! Når du har krysset av et svar kan du ikke endre det til et blankt svar etterpå. Du kan gå frem og tilbake i oppgavesettet og justere svaret ditt. (Hvis du ikke har krysset av kan du derimot gå tilbake og sette kryss senere).

Håndregning
Der det er angitt skal du også levere inn håndregning. Håndregning gjør du på eget ark, tar bilde og laster opp i Inspera. Tips til hvordan gjøre dette enkelt er gitt i emnets side i Canvas. Selv om endelig svar skal oppgis i eget felt vil du **ikke få poeng** uten at håndregningen er vedlagt. Dette gjelder for alle oppgaver der håndregning skal lastes opp.

Det vil være andre oppgaver der det kreves noe regning, men du skal bare laste opp bilde i de oppgavene der det står tydelig at det skal gjøres.

Bruk av software
I noen oppgaver skal du bruke software til å analsyere data. Du må ha statistikksoftware installert på egen datamaskin. For å få uttelling på disse oppgavene skal full kode limes inn som angitt i oppgaveteksten.

Av tekniske grunner vil Inspera endre navnet på datafilen når den lastes ned, ved å legge til en del siffer på slutten av filnavnet. Du må derfor være nøye på å gi datafilen riktig navn i scriptet slik at det leses inn riktig. (Du kan evt. endre filnavnet før du åpner filen, men det trengs ikke).

For eksempel vil et datasett i denne oppgavesettet ha navnet HR_analytics, men når det lastes ned til download-mappen din har filen fått navnet HR_analytics_rd_81018196_1619593996048. Det vil heller ikke være en filhale .rds. Hvis du kopierer filen fra download-mappen til data-mappen din vil du kunne lese det inn på vanlig måte med følgende: readRDS("data/HR_analytics_rd_81018196_1619593996048")

Om poenggivning

For alle typer oppgaver får du 1 poeng per riktige svar hvis ikke annet er angitt.

For alle spørsmål der du skal krysse av for riktig svar (dvs. flervalgsoppgave) får du 1 poeng hvis riktig, men blir trukket 0.25 poeng for hvert feil svar. Ubesvarte gir null poeng. Med andre ord: Du bør altså ikke gjette hvis du ikke er ganske sikker på svaret.

I andre oppgaver skal du skrive inn riktig tall eller skrive noe tekst. I disse oppgavene får du ikke trukket poeng dersom du svarer feil.

For oppgavetyper der du skal fylle inn riktig tall er det viktig at du svarer med riktig antall desimaler. Galt antall desimaler kan medføre at svaret blir regnet som feil.

Merk: Du kan gå frem og tilbake i oppgavesettet og justere svarene gjennom hele eksamenstiden.
OBS! Om du avgir svar i en flervalgsoppgave kan du ikke fjerne svaret ditt og levere ubesvart. Vær derfor helt sikker på at du ønsker å avgi svar før du krysser av.

Innlevering i Inspera

- Les mer om eksamen i Inspera.
- Når du har begynt på eksamen / lastet opp din besvarelse, vil du se at besvarelsen er lagret.
- Du kan endre besvarelsen din frem til innleveringsfristen.
- Når innleveringsfristen går ut, leveres besvarelsen automatisk.
- Hvis du etter å ha begynt på eksamen likevel ikke vil levere besvarelsen, må du trekke deg fra eksamen. Trykk på ikonet øverst i høyre hjørne for å få opp valget "Jeg vil trekke meg".

Spørsmål under eksamen

- Hvis du har spørsmål under eksamen, må du sende e-post fra din UIO-mail til hjemmeeksamen@sv.uio.no. Husk å skrive emnekoden i emnefeltet.
- Hvis vi må gi informasjon under eksamen, blir det publisert i Canvas. Sørg for å ha skrudd på varselinnstillinger i Canvas.

Etter eksamen

- Etter innleveringsfristen må du sjekke at din innleverte besvarelse ligger i Inspera under **Arkiv.**
- Send e-post til hjemmeeksamen@sv.uio.no om du ikke ser besvarelsen. Legg ved besvarelsen hvis du har den som en fil.
- *Flersvalgsspørsmålene på eksamen skal kunne gjenbrukes, og vil derfor vil ikke eksamensoppgaven din bli tilgjengelig i etterkant av eksamen.*

12 I en ensidig test av ett gjennomsnitt får du oppgitt følgende:
df=14 og t=2.3. Hva er da p-verdien? Bruk tabellen som gir det mest nøyaktige svaret og oppgi svaret med **to desimaler** her: [ ] (0.02).

Maks poeng: 1

13 En venn av deg hevder at Oslo kommune sin satsning på skolen har ført til ønskede resultater og dermed at elever i Oslo gjør det bedre enn f.eks. elever fra Bergen. For å undersøke om det kan stemme får dere tak i noe data.

Oslo og Bergen er de to byene med flest avgangselever fra ungdomsskolen. I tabellen under finner du informasjon om grunnskolepoengene til et tilfeldig utvalg av elever fra Oslo og Bergen for et gitt kull.

	Antall elever	Gjennomsnittlig grunnskolepoeng	Standardavviket til gjennomsnittet
Oslo	2735	44,45	11,15
Bergen	2040	44,32	11,08

Er det slik at grunnskolepoeng er systematisk lavere for avgangselever i Bergen sammenlignet med avgangselever i Oslo for dette kullet? Eller kan det skyldes tilfeldig variasjon?

OBS! Gjør hele utregningen og last opp bilde av utregningen for oppgaven på neste side i oppgavesettet. Utregning må ligge vedlagt for å få uttelling på denne oppgaven.

Oppgave a)
Hvorfor skal du her gjøre en ensidig test?

Velg ett alternativ

- Utgangspunktet er at man tester om Oslo har høyere snitt enn Bergen eller ikke. O[illegible] Bergen skulle ha høyere snitt enn Oslo spiller ingen rolle. ✔
- Nullhypotesen er om byene er forskjellige eller ikke, og det er da bare en alternativ hypotese
- Du er ikke fra Bergen selv
- Nullhypotesen er at Oslo har høyere gjennomsnitt enn Bergen

Gjør en en-halet (også kalt ensidig) t-test for om det er en forskjell i grunnskolepoeng for elever fra Oslo og Bergen i dette avgangskullet ved å løse oppgave a)-e) under. Bruk gjennomgående **to desimaler** i utregningen.

Oppgave b)
Hva er differansen mellom gjennomsnittet for Oslo og gjennomsnittet for Bergen? Skriv svaret ditt her (bruk **to desimaler** i svaret): [ ] (0.11 - 0.15)

Oppgave c)

Hva er standardfeilen til differansen mellom gjennomsnittet for Oslo og gjennomsnittet for Bergen? Skriv svaret ditt her (bruk **to desimaler** i svaret): [ ] (0.325 - 0.335)

Oppgave d)

Hva er resultatet av t-testen (dvs. t-verdien?) Skriv svaret ditt her (bruk **to desimaler** i svaret):

(0.37 - 0.41)

Oppgave e)

Hva er intervallet for p-verdi?

Velg ett alternativ

- 0.02 > p > 0.01
- 0.0025 > p > 0.001
- 0.025 > p > 0.02
- 0.20 > p > 0.15
- 0.01 > p > 0.005
- 0.25 > p>0.20
- p < 0.0005
- 0.15 > p > 0.10
- 0.005 > p > 0.0025
- 0.10 > p > 0.05
- 0.001 > p > 0.0005
- 0.05 > p > 0.025
- p > 0.25 ✔

Oppgave f)

Hva er konklusjonen på testen, ved et 5 % konfidensnivå?

Velg ett alternativ

- Grunnskolepoeng i Bergen er lavere enn i Oslo i større grad enn hva vi ville forvente fra tilfeldig variasjon (dvs. statistisk signifikant)
- Grunnskolepoeng i Bergen er lavere enn i Oslo, men ikke mer enn hva vi kunne forvente fra tilfeldig variasjon (dvs. ikke statistisk signifikant) ✔

Antall elever
Gjennomsnittlig grunnskolepoeng
Standardavviket til gjennomsnittet
Oslo
2735
44,45
11,15
Bergen
2040
44,32
11,08

Maks poeng: 3

14 Last opp bilde av håndregningene du gjorde til deloppgave b)-d) på forrige oppgave (Grunnskolepoeng) her.

Last opp filen her. Maks én fil.

Alle filtyper er tillatt. Maksimal filstørrelse er **2 GB**.

Velg fil for opplasting

Maks poeng: 0

15 I en artikkel på NRK fra 2019 hevdes det at ny forskning viser at selv moderat fysisk aktivitet kan redusere risikoen for tidlig død. Artikkelen viser til en såkalt metastudie som sammenlikner dødelighet blant mer enn 36.000 personer over 40 år som har oppgitt hvor mange minutter de beveger seg daglig. Artikkelen fremhever blant annet at de 15 minuttene som skiller den «slappeste» og nest «slappeste gruppen» henger sammen med en 25% lavere risiko for tidlig død, og «rolige gåturer, litt hagearbeid og det å tusle til butikken eller å gå til puben gjør faktisk en stor forskjell for veldig mange av oss».

Det er flere metodiske svakheter knyttet til artikkelens argumenter for at det å øke den fysiske aktiviteten har en kausal effekt på dødeligheten.

Oppgave a)
Beskriv **den mest sentrale svakheten** ved studiens ambisjon om å si noe om effekten av fysisk aktivitet på dødelighet og gi konkrete eksempler.

Skriv ditt svar her

Oppgave b)
Gi ett eksempel på hvordan analysen kunne blitt **forbedret**.

Skriv ditt svar her

Maks poeng: 3

16 I løpet av Korona-utbruddet i Norge har mediene skrevet mye om at smittetallene er særlig høye i innvandrertette bydeler i Oslo, og at deler av årsaken kan skyldes at smitteverntiltakene av ulike årsaker ikke følges like godt der som i andre deler av byen. Du snakker med en kompis av deg som har lyst til å undersøke dette nærmere på egenhånd. Planen hans er å stille seg utenfor to kjøpesentre i ulike deler av byen og undersøke hvor stor andel av kundene som bruker munnbind på vei inn på senteret, og deretter sammenligne tallene for å si noe om i hvilken grad personer fra bydeler med ulik innvandrertetthet overholder smitteverntiltakene. Han velger seg ut kjøpesenteret CC Vest i bydel Ullern på «Vestkanten» og Tveita Senter i bydel Alna på «Østkanten», og planlegger å stå utenfor hovedinngangen to onsdager på rad mellom kl. 12 og kl. 13.

Kompisen din vet at du akkurat har tatt SOSGEO1120 på Blindern, og lurer på hva du synes om forskningsdesignet hans. Fortell ham om inntil to sentrale utfordringer med designet hans slik han har skissert det.

Skriv ditt svar her

Maks poeng: 2

17

Under er det flere oppgaver tilknyttet de viste tabellene. Vær obs på at oppgave d) og e) krever litt håndregning. *Bilde av håndregningen lastes opp i én fil på neste side i oppgavesettet.*

Tabell 1 gir resultatene fra en regresjonsanalyse av sammenhengen mellom det å ha innvandringsbakgrunn og regneferdigheter (målt ved poeng på nasjonale prøver).

Vær obs på at det er benyttet engelsk notasjon (som i pensum) der *punktum* er desimaltegn.

Tabell 1: Regresjonsanalyse av sammenhengen mellom innvandringsbakgrunn og regneferdigheter.

	Nasjonale prøver: Regning
Innvandringsbakgrunn	-2.455***
	(0.148)
Kvinne	-1.676***
	(0.0552)
Foreledreinntekt, rangert	0.116***
	(0.00109)
Innvandringsbakgrunn x Foreledreinntekt, rangert	0.0446***
	(0.00407)
Konstantledd	49.85***
	(11.39)
N	170560

Data tilgjengelig gjennom prosjektet «Ethnic segregation in schools and neighborhoods: consequences and dynamics», finansiert av Forskingsrådet (prosjektnummer 236793).
Alle modeller inkluderer dummyer for avgangsår fra grunnskolen.
Standardfeil i parentes.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Oppgave A:

Gitt denne modellen, hvilket av følgende utsagn stemmer?

Blant gutter med lavest foreldreinntekt...

Velg ett alternativ

- ...har de med innvandringsbakgrunn gjennomsnittlig høyere regneferdigheter enn de uten innvandringsbakgrunn
- ...har de med innvandringsbakgrunn gjennomsnittlig lavere regneferdigheter enn de uten innvandringsbakgrunn ✔
- ...er det i gjennomsnitt ingen forskjell mellom de med og uten innvandringsbakgrunn når det kommer til regneferdigheter

Oppgave B:

Du skal, ut fra denne modellen, vurdere hvilken rolle foreldreinntekt spiller for elevenes leseferdigheter. Hvilket av følgende utsagn er riktig?

Velg ett alternativ

- Modellen er uegnet til å fortelle oss noe om sammenhengen mellom foreldrenes inntekt og regneferdigheter
- Elever med høyere foreldreinntekt ser ut til å ha høyere regneferdigheter ✔
- Det er en negativ sammenheng mellom foreldrenes inntekt og regneferdigheter
- Det er ingen signifikant sammenheng mellom foreldrenes inntekt og regneferdigheter
- Foreldreinntekt øker som følge av økte regneferdigheter hos eleven

Oppgave C:

Det kan være at foreldreinntekt spiller ulik rolle for elever med og uten innvandringsbakgrunn. Gitt modellen, hvilket av følgende er riktig?

Velg ett alternativ

- Det er ingen signifikant forskjell i regneferdigheter når vi sammenlikner elever med og uten innvandringsbakgrunn som har samme foreldreinntekt
- Foreldreinntekt betyr mer for regneferdigheter til elever med innvandringsbakgrunn enn for regneferdigheter til elever uten innvandringsbakgrunn ✔
- Foreldreinntekt betyr 0 for regneferdigheter til elever uten innvandringsbakgrunn

Oppgave D:

Regn ut forskjellen i poeng på nasjonale prøver i lesing for elever med innvandringsbakgrunn som har foreldreinntekt lik 30 sammenliknet med foreldreinntekt lik 31. Bruk fire desimaler i utregningen.

Vis hele utregningen for hånd og last opp bilde på neste side i oppgavesettet.

Skriv svaret her med fire desimaler: (0.561 - 0.563)

Oppgave E:
Du skal predikere leseferdigheter for jenter med innvandringsbakgrunn som har foreldre med inntektsrangering lik 50.

Regn ut predikerte poeng på nasjonale prøver i lesing for denne gruppen av elever. Bruk alle oppgitte desimaler i utregningen.

Vis hele utregningen for hånd og last opp bilde på neste side i oppgavesettet.

Skriv svaret med fire desimaler her: (53.746 - 53.752)

Maks poeng: 5

18 Last opp bilde av håndregningene du gjorde til deloppgave d) og e) på forrige oppgave (regresjon) her.

Last opp filen her. Maks én fil.

Alle filtyper er tillatt. Maksimal filstørrelse er **2 GB**.

Velg fil for opplasting

Maks poeng: 0

19 Vedlagt er et datasett i csv-format. Last det ned fra denne lenken: datasett og les det inn i R (evt. Stata).
Datasettet har to variable med navn x og y. Lag et scatterplot med x-variabelen langs x-aksen og y-variabelen langs y-aksen.

Hva viser plottet?
Riktig svar gir 2 poeng.

Velg ett alternativ

- Donald Trump
- Et skrått kryss (dvs. X)
- En stjerne
- Donald Duck ✔
- En dinosaur
- Ikke noe tydelig mønster

Maks poeng: 2

20 For ethvert firma er gjennomtrekk av ansatte en utfordring bl.a. fordi de mister kompetanse som er bygd opp over tid. Det kan derfor være en fordel om man kan forhindre gjennomtrekk.

I et stort firma har personalavdelingen samlet gjort en spørreundersøkelse blant de ansatte, og noen år senere har de koblet denne informasjonen til om vedkommende har sluttet i jobben eller ikke. Dette kan ha betydning for både planlegging og fremtidige ansettelser.

Alle oppgavene nedenfor skal løses med R eller annen statistikksoftware. Fullt script skal limes inn på neste side i oppgavesettet for å få uttelling.

Datasettet laster du ned i RDS-format her: HR_analytics (Evt. som csv-format her: HR_analytics_csv)

Variablene i datasettet er som følger:
slutter - en dummy for om personen slutter i jobben eller ikke (0 = ikke slutter, 1 = slutter)
distancefromhome - reisevei til hjemmet, oppgitt i miles
male - dummy for kjønn (0 = kvinne, 1 = mann)
age_lt_30 - dummy for å være under 30 år (1 = yngre enn 30 år, 0 = 30 år eller eldre)
jobsatisfaction - oppgitt tilfredshet med jobben. Kontinuerlig variabel, høy verdi er mer tilfreds
worklifebalance - balanse mellom jobb og fritid. Fire kategorier: Bad, Good, Better, Best
overtime - dummy for å jobbe mye overtid (0 = nei, 1 = ja)

Oppgave a)

Hvor stor andel av de ansatte har sluttet i perioden? Skriv svaret her med 3 desimaler:

(0.160 - 0.169)

Oppgave b)
Personalsjefen har en hypotese om at en viktig grunn til at folk slutter i jobben er problemer med å balansere tidsbruk mellom jobb og fritid. Lag en tabell som viser andel som slutter fordelt etter variabelen worklifebalance. I tabellen er det særlig en gruppe som skiller seg ut med høy andel som slutter i jobben. Hvor stor andel slutter i denne kategorien?

Skriv inn svaret her med 3 desimaler her: (0.310 - 0.319)

Oppgave c)
Selv om det er forskjeller i andel som slutter kan det likevel skyldes tilfeldigheter. Undersøk om den observerte sammenhengen mellom worklifebalance og å slutte i jobben kan sies å være systematisk eller om det kan skyldes tilfeldig variasjon. Bruk den testen som er best egnet. Hva er p-verdien for denne testen?

Skriv inn svaret med fem desimaler her: (0.00096 - 0.00098)

Oppgave d)
Hva er konklusjonen på denne testen?

Velg ett alternativ

- Det er en tydelig årsakssammenheng mellom worklifebalance og slutte i jobben (statistisk signifikant)
- Den observerte sammenhengen mellom worklifebalance og slutte i jobben er ikke større enn at det kan skyldes tilfeldig variasjon i dataene
- Den observerte sammenhengen mellom worklifebalance og slutte i jobben er lite sannsynlig at skyldes tilfeldig variasjon i dataene ✔
- Det er en tydelig at å slutte i jobben forårsaker dårlig worklifebalance (statistisk signifikant)
- Den observerte sammenhengen mellom worklifebalance og slutte i jobben kan helt sikkert ikke forklares med tilfeldig variasjon i dataene

Oppgave e)

Det er en rekke andre faktorer som man antar spiller inn på om man vil slutte i jobben, som f.eks. trivsel, reisevei til jobben, kjønn og alder. Det kan derfor være lurt å kontrollere for disse variablene når man ser på sammenhengen med worklifebalance.

Estimer en multippel regresjonsmodell der du ser på sannsynligheten for å slutte. Hvordan vil du beskrive sammenhengen mellom worklifebalance og slutte i jobben når du kontrollerer for de nevnte variablene?

Velg ett alternativ

- Bare de med worklifebalance "Better" har en statistisk signifikant lavere sannsynlighet for å slutte i jobben
- Alle med worklifebalance bedre enn "bad" har lavere sannsynlighet for å slutte i jobben. Men forskjellen er ikke statistisk signifikant
- De andre variablene det kontrolleres for forklarer sammenhengen slik at worklifebalance ikke trengs
- Alle med worklifebalance bedre enn "bad" har lavere sannsynlighet for å slutte i jobben. Disse forskjellene er statistisk signifikante på 5% nivå. ✔
- Fordi R-squared er lav og residual standard error er høy, gir det ikke mening å tolke estimatene for worklifebalance

Oppgave f)

Fra regresjonsmodellen i forrige deloppgave er det tydeligvis to andre grupper som oftere slutter i jobben: yngre arbeidstakere og de som jobber mye overtid.

I følge denne modellen, for to personer som ellers har like kjennetegn, hva er forskjellen i sannsynlighet for at en som jobber mye overtid slutter sammenlignet med en som ikke jobber mye overtid?

Skriv svaret her med 3 desimaler: (0.200 - 0.205)

Oppgave g)

Du har så en hypotese om at det er særlig kombinasjonen av å være yngre arbeidstaker *og* jobbe mye overtid som øker risikoen for å slutte. Estimer en regresjonsmodell for å teste denne påstanden.

Hva er den estimerte regresjonsparameteren som belyser denne hypotesen?

Skriv svaret her med 3 desimaler: (0.220 - 0.229)

Oppgave h)

Basert på svaret i forrige oppgave, hva er en rimelig konklusjon?

Velg ett alternativ

- Å jobbe mye overtid øker sannsynligheten for å slutte, men i større grad for de under 30 enn de over 30 år ✔
- Å jobbe overtid reduserer sannsynligheten for å slutte, men ikke for de over 30 år
- Det er alder og ikke overtid som påvirker sannsynligheten for å slutte
- Å jobbe mye overtid øker sannsynligheten for å slutte, men det er ingen statistisk signifikant forskjell mellom de under og de over 30 år

Kilde: Datasettet "IBM HR Analytics Employee Attrition & Performance" er tilgjengelig via Kaggle: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

---

Maks poeng: 8

21 Scriptet til løsningen i forrige oppgave om HR-analytics limer du inn i feltet nedenfor.

Lim inn koden her:

```
1
```

Maks poeng: 0